Representation and Analysis of Signals. Part XXVI. Least-Squares Approximation of Functions by Exponentials.

JOHNS HOPKINS UNIV BALTIMORE MD

AUG 1969

...sified
Security Classification

DOCUMENT CONTROL DATA - R&D

...ecurity classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

...GINATING ACTIVITY (Corporate author)
...epartment of Electrical Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
Least-Squares Approximation of Functions by Exponentials

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (Last name, first name, initial)
Miller, Gerry

6. REPORT DATE
August 1969

7a. TOTAL NO. OF PAGES
84

7b. NO. OF REFS
36

8a. CONTRACT OR GRANT NO.
Nonr-4010(13)

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
XXVI

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
None

10. AVAILABILITY/LIMITATION NOTICES
Qualified requesters may obtain copies of this report from DDC.

11. SUPPLEMENTARY NOTES Distributed to recipients under contract Nonr-4010(13) with permission of the author.

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research
Code 468, Washington, D.C.

13. ABSTRACT

The approximation of an analytic time function in the least squares sense by sums of exponentials is considered from several different points of view. In particular, we consider the determination of the 2n complex parameters $\{\alpha_k, s_k\}$ of the function $f_a(t) = \sum_{k=1}^{n} \alpha_k \exp(s_k t)$ so that for a given n and f(t), the value of the functional

$$J = \int_0^\infty [f(t) - f_a(t)]^2 \, dt$$

is minimized. Equivalently, the 2n real parameters $\{a_k, b_k\}$, of the Laplace transform

$$F_a(s) = \frac{a_1 + a_2 s + \ldots + a_n s^{n-1}}{b_1 + b_2 s + \ldots + b_n s^{n-1} + s^n}$$

of $f_a(t)$, may be determined to achieve the same minimum value of

Unclassified
Security Classification

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	W
Least-Squares Estimation Exponential Representation Orthogonal Exponentials Approximation						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional

Security Classification

13. ABSTRACT (Continued)

error, J.

McDonough's method for finding the $\{s_k\}$ is derived by three different approaches. In the process, a new method is developed which offers the advantages of the earlier results achieved by McDonough and by McBride, Schaefgen, and Steiglitz. This new method reveals the link between these earlier methods and provides a standard for comparing these two linear iterative schemes using several numerical examples.

The linear least-squares approximation procedure in which both n and the $\{s_k\}$ (or $\{b_k\}$) are fixed is also discussed in detail. Examples show the numerical difficulties due to roundoff errors that arise even with the straightforward methods available to find the $\{\alpha_k\}$. A simple criterion for estimating these errors before finding the $\{\alpha_k\}$ is developed to permit one to evaluate the feasibility of obtaining accurate results in any given situation.

June 1969

THE JOHNS HOPKINS UNIVERSITY
DEPARTMENT OF ELECTRICAL ENGINEERING
BALTIMORE, MARYLAND 21218

REPRESENTATION AND ANALYSIS OF SIGNALS

PART XXVI. LEAST-SQUARES APPROXIMATION OF FUNCTIONS BY EXPONENTIALS

by

Gerry Miller

A report distributed under
Contract Nonr-4010(13) with
the Office of Naval Research.

LEAST-SQUARES APPROXIMATION OF FUNCTIONS BY EXPONENTIALS

ABSTRACT

The approximation of an analytic time function in the least-squares sense by sums of exponentials is considered from several different points of view. In particular, we consider the determination of the 2n complex parameters $\{\alpha_k, s_k\}$ of the function $f_a(t) = \sum_{k=1}^{n} \alpha_k \exp(s_k t)$ so that for a given n and f(t), the value of the functional

$$J = \int_0^\infty [f(t) - f_a(t)]^2 \, dt$$

is minimized. Equivalently, the 2n real parameters $\{a_k, b_k\}$, of the Laplace transform

$$F_a(s) = \frac{a_1 + a_2 s + \ldots + a_n s^{n-1}}{b_1 + b_2 s + \ldots + b_n s^{n-1} + s^n}$$

of $f_a(t)$, may be determined to achieve the same minimum value of error, J.

McDonough's method for finding the $\{s_k\}$ is derived by three different approaches. In the process, a new method is developed which offers the advantages of the earlier results achieved by McDonough and by McBride, Schaefgen, and Steiglitz. This new method reveals the link between these earlier methods and provides a standard for comparing these two linear iterative schemes using several numerical examples.

The linear least-squares approximation procedure in which both n and the $\{s_k\}$ (or $\{b_k\}$) are fixed is also discussed in detail. Examples show the numerical difficulties due to roundoff errors that arise even with the straightforward methods available to find the $\{\alpha_k\}$. A simple criterion for estimating these errors before finding the $\{\alpha_k\}$ is developed to permit one to evaluate the feasibility of obtaining accurate results in any given situation.

ACKNOWLEDGEMENTS

I have been most fortunate to have had the opportunity to work under the direction of William H. Huggins, Westinghouse Professor of Electrical Engineering. This thesis is to a large extent the result of his patient and inspiring guidance.

I would also like to take this opportunity to thank Dr. Stephen Wolff for reading and commenting on the entire manuscript and who, along with Dr. Huggins, served as referee. I am indebted to fellow graduate students, in particular Kenneth Lutz and Richard Healy, with whom I discussed my research on several occasions, and to Mrs. R. Howard for the excellent typing.

Finally, I am grateful to my wife, Barbara, for her patience and confidence in me.

TABLE OF CONTENTS

Table of Contents (cont'd.)

LEAST-SQUARES APPROXIMATION OF FUNCTIONS BY EXPONENTIALS

I. INTRODUCTION

I.1 Exponential Representations

In approximating a function of time, such as might arise in control or communication problems, one commonly uses a linear combination of a finite set of simpler functions. Exponential functions are particularly appropriate for this purpose because they have simple mathematical properties. It has been demonstrated in [1]-[4][†] that exponentials have very good approximation properties for a rather broad range of signal wave shapes. Furthermore, in linear time-invariant systems, the class of exponential functions provides a natural representation since the natural response of these systems is composed of exponential components. Another feature of an exponential representation is that there are arbitrarily many different discrete sets $\{\exp(s_k t)\}$ that are complete over the semi-infinite interval with respect to the L_2 norm (i.e. the mean-square error in approximating any piecewise continuous $f(t)$ that is square integrable over $0 < t < \infty$ by the form $\sum_k \alpha_k \exp(s_k t)$ can be made arbitrarily small). This completeness property is established by Szasz's form of Muntz's theorem [5], which when applied to this exponential basis may be stated as follows: The basis $\{\exp(s_k t)\}$ is fundamental with respect to the L_2 norm over the semi-infinite interval if and only if

$$\sum_{k=1}^{\infty} - \frac{\mathrm{Re}(s_k)}{1+|s_k+\frac{1}{2}|^2} \Rightarrow \infty. \qquad (1.1)$$

[†] Whole numbers in brackets refer to references listed beginning on page 81.

However, for practical work we are not interested in letting k approach infinity. Instead, we seek efficient representation in which k is small. Of course, any finite representation will necessarily be approximate and incomplete. We are interested in finding the basis of lowest possible dimension that will lead to an approximation of acceptable accuracy. Efficient representation will enable us to extract the information-bearing attributes of the signal with a minimum of processing. When the interval of approximation is finite, one can resort to the discrete Fourier series since sines and cosines belong to the class of the exponential functions. But, despite the popularity of Fourier representation, one can often do better than this for pulse-like signals by using more general exponential components. For this reason exponential functions play an important role in signal representation.

To best approximate a signal by a set of n exponentials, one must determine the n "optimum" exponents s_k as well as the n amplitudes α_k. These exponents and amplitudes may be chosen to minimize the error with respect to some norm. Two popular norms are the integrated squared error (L_2 norm)

$$J = \int_0^\infty [f(t) - \sum_{k=1}^{n} \alpha_k \exp(s_k t)]^2 dt = \int_0^\infty e^2(t)\, dt \qquad (1.2)$$

and the Chebyshev norm (uniform norm), $\max_{t>0} |e(t)|$. The former, often referred to as the least-squares (or minimum-error energy) criterion has been studied extensively because it is the most tractable mathematically. For given $\{s_k\}$, it is easy in principle to choose the $\{\alpha_k\}$ for the least-squares criterion, since $f_a(t)$ is a linear function of the $\{\alpha_k\}$. However, practical computational difficulties exist because the exponential functions are highly correlated. As a consequence, solutions of the $\{\alpha_k\}$ may be subject to large errors due to roundoff in the numerical computation. This difficulty will be examined further in chapter II.

Difficulty of a more serious nature arises in finding the exponents $\{s_k\}$ for a given f(t) that satisfy the minimum error energy criterion. Until recently, only gradient methods were available, and these frequently proved to be quite unwieldly for large n. Then in 1966 McBride, Steiglitz, and Schaefgen [6] and in 1968 McDonough and Huggins [7] developed two different linear iterative schemes which have been found to be quite successful for determining the $\{s_k\}$ even for large n. Two natural questions about these methods are the following. First, how are these methods related? Second, when is it advantageous to use one method rather than the other? This thesis provides answers to these questions by developing a new linear iterative method under the least-squares criterion.

The Chebyshev or uniform-norm criterion has been studied less than the least-squares criterion because it is analytically more difficult. Apparently, not much has been done with this criterion to date but, Tang [8] has shown how the $\{\alpha_k\}$ may be determined provided all the $\{s_k\}$ are real and distinct. So far, it appears that the only way to find the exponents $\{s_k\}$ for the Chebyshev criterion is by slowly converging gradient methods.

I.2 Some Known Methods of Approximation by Exponentials

I.2.1 Non-Optimal Approximation - Prony's Method and Padé Approximants

Two simple, but often successful ways of obtaining an approximation to a function by sums of exponentials use Prony's method and Padé approximants. Neither results in an optimal approximation with respect to the L_2, uniform, or any other norm, but they do provide two quick and straightforward ways of obtaining approximations that are usually "fairly good". In Padé approximants one matches the rational function

$$F_a(s) = \frac{\sum_{k=1}^{m} a_k s^{k-1}}{\sum_{k=1}^{n+1} b_k s^{k-1}} = \frac{M(s)}{D(s)}, \quad b_{n+1} = 1 \tag{1.3}$$

to the desired function F(s) (the Laplace transform of f(t)) by adjusting the $\{a_k, b_k\}$ such that $F_a(s)$ will have the same power series as the power series expansion of F(s) through the m+nth power where $m \leq n$. That is, the seminorm

$$||F(s)-F_a(s)||=|F(0)-F_a(0)|+|F'(0)-F_a'(0)|+\ldots+|F^{m+n}(0)-F_a^{m+n}(0)| \tag{1.4}$$

is made zero. The real merit of the Padé method is the computational ease with which the $\{a_k, b_k\}$ may be found. Finding the $\{b_k\}$ involves solving n linear equations in n unknowns. Once the $\{b_k\}$ are determined, the $\{a_k\}$ are similarly found by evaluating another linear system of m equations. These are explicit equations, not simultaneous for the $\{a_k\}$.

To write $F_a(s)$ as a sum of exponentials, a partial-fraction expansion must be performed which requires finding the roots of the nth degree polynomial D(s). Kautz [9] and Mathers [10] have used the method in designing circuits to approximate specified transient responses. Teasdale [11] first applies the transformation $z=(1-s)/(1+s)$ to obtain an "indirect Padé approximant" matching a power series in z instead of s. (Actually, since z=0 implies s=1, this is matching the power series about the point s=1 instead of s=0.) The procedure developed will be different from the direct Padé approximant with generally smaller error but at the expense of more computation.

Another simple way of approximating a function by sums of exponentials is a method first used by Prony in 1795. This procedure was first applied to network synthesis problems by Tuttle, Carr, and Kautz. A detailed discussion of the method and its refinements is given in McDonough's thesis [1]. The principle behind the method originates from the fact that if a waveform is indeed composed of exponentials, viz.

$$f(t) = \sum_{k=1}^{n} \alpha_k \exp(s_k t) \qquad \mathrm{Re}\{s_k\} < 0 \tag{1.5}$$

then f(t) will be the solution to some homogeneous differential equation of the n^{th} order,

$$\sum_{i=o}^{n} B_i \frac{d^i f}{dt^i} = 0 \quad , B_o = 1 \tag{1.6}$$

Provided one could find the coefficients $\{B_i\}$ of this equation, the $\{s_k\}$ could then be obtained by evaluating the n roots of the polynominal $\sum_{i=o}^{n} B_i s^i = 0$. Our task then is to find the B_i appropriate to a given f(t). Then, the $\{s_k\}$ which satisfy the differential equation may be used to construct an exponential basis for f(t). If the signal is noisy or is not exactly the sum of n exponentials, the left hand side of equation (1.6) cannot be made zero regardless of the choice of $\{B_i\}$ and there will be a residual $\sum_{i=o}^{n} B_i(d^i f/dt^i) = \epsilon_p(t)$.

Since $B_0=1$, equation (1.6) may be written as

$$\epsilon_p(t) = f(t) + \sum_{k=1}^{n} B_k \frac{d^k f(t)}{dt^k} \qquad (1.7)$$

Then, one simply chooses the set of $\{B_i\}$ to minimize this $\epsilon_p(t)$ in the least-squares sense, thus

$$E = \int_0^\infty [\epsilon_p(t)]^2 \, dt. \qquad (1.8)$$

Minimizing E over the coefficients $B_1, B_2, \ldots, B_n$ results in n linear simultaneous equations

$$\frac{\partial E}{\partial B_i} = \int_0^\infty f^{(i)} f \, dt + \sum_{k=1}^{n} B_k \int_0^\infty f^{(i)} \, f^{(k)} \, dt = 0 \qquad (1.9)$$

$$i = 1, 2, \ldots, n.$$

However, the matrix elements

$$g_{ik} = \int_0^\infty f^{(i)} \, f^{(k)} \, dt \qquad (1.10)$$

will not exist unless f is of at least class C^n. If the differential equation is first integrated n times the corresponding new elements will exist for any piecewise continuous function with finite energy but, this initial integration should be performed the least number of times to assure the existence of (1.10) since it has a tendency to destroy signal information. Fortunately, the matrix elements g_{ik} have certain recursion relations which make it necessary to calculate the g_{kk} only. Prony's method yields only the frequencies $\{s_k\}$ but, the amplitude coefficients $\{\alpha_k\}$ may subsequently be found with little difficulty (as discussed in the next chapter). It should be emphasized that Prony's method does not lead to the optimum least-square approximation (unless $f(t)$ is exactly the sum of n exponentials) since $\varepsilon_p(t)$ is not identical to $f(t)-f_a(t)$.

I.2.2 Optimal Approximation in the Least-Squares Sense by Exponentials

The conditions for optimal exponential approximation of a function $f(t)$ with respect to the L_2 norm over the semi-infinite interval are described compactly by the equations of Aigrain and Williams [12]. Although theoretically attractive, these nonlinear transendental equations are computationally undesirable and algebraic solution is seldom possible even when the Laplace transform of $f(t)$ is known in closed form. Most often, these equations are solved by a gradient or some other iterative method.

In chapter III, it will be shown that by suppressing the amplitude coefficients $\{\alpha_k\}$ one may write the integrated square error, J, as

$$J = \int_0^\infty f^2(t)\,dt - \underline{\tilde{F}}\,\underline{\underline{H}}^{-1}\,\underline{F} \tag{1.11}$$

where $\underline{\underline{H}}^{-1}$ is the inverse of the generalized Hilbert matrix and $\underline{F}$ is a column matrix ($\underline{\tilde{F}}=[F(s_1^*),F(s_2^*),\ldots,F(s_n^*)]$). Equation (1.11) is a compact mathematical expression for the mean-square error but, for large n, (say $n\geq5$) it is very sensitive to the variation of the $\{s_k\}$ and finding the minimum J by the usual gradient methods may be ineffective. The two much more effective ways of solving the Aigrain-Williams equations for large n, which have been recently developed, shall now be briefly discussed.

The method of McBride, Schaefgen, and Steiglitz, the first of the linear iterative methods mentioned earlier, introduces an approximate error

$$E_a(s) = \frac{D_j(s)}{D_{j-1}(s)} F(s) - \frac{N_j(s)}{D_{j-1}(s)} \qquad (1.12)$$

with

$$F_a(s) = \frac{N(s)}{D(s)} = \frac{a_1+a_2s+\ldots+a_ns^{n-1}}{b_1+b_2s+\ldots+b_ns^{n-1}+s^n} \qquad (1.13)$$

where the subscript j refers to the iteration number. The previously computed coefficients of D_{j-1} are regarded as fixed during the j^{th} iteration. By this simple tactic, the error is linearized in terms of the unknown coefficients $\{a_k,b_k\}$ of the numerator polynomials N_j and D_j. The primary difficulty with this method is that the approximate rather than the true error is being minimized. Hence, the iterative scheme does not converge to the true optimum. To circumvent this difficulty, McBride et. al. introduce a different "Mode-2 Iteration" which does converge to the true minimum but more slowly than one would hope. The requirement for using two different iteration schemes also adds extra complexity to the McBride method.

The difficulties of the McBride method are avoided in the linear iterative scheme devised by McDonough and Huggins. Here, the 2n Aigrain-Williams equations are first reduced to a set of n equations involving the $\{s_k\}$ only. This was done by regarding $F(s)$ as a signal in a vector space and showing that a necessary condition for the $\{s_k\}$ to be optimum is that $F(s)$ be orthogonal to the space spanned by $\phi_i(s)$ $i = 1,2,\ldots,n$ where

$$\phi_i(s) = \frac{1}{\sqrt{-s_i-s_i^*}} \frac{H(s)}{s-s_i} \prod_{k=1}^{i-1} \frac{(s+s_k)}{(s-s_k)} \qquad (1.14)$$

with

$$H(r) = \prod_{k=1}^{n} \frac{(s+s_k)}{(s-s_k)} = (-1)^n \frac{D(-s)}{D(s)} . \qquad (1.15)$$

This orthogonality condition may be written as

$$\int_{-j\infty}^{j\infty} F(-s)\, \phi_i(s) \frac{ds}{2\pi j} = 0 \qquad i = 1,2,\ldots,n. \qquad (1.16)$$

The linear iterative scheme described by McDonough is obtained by replacing H(s) with the new operator

$$H_a(s) = \sum_{i=1}^{n+1} b_i(-s_i)^{i-1} \;/\; D(s) \qquad , b_{n+1} = 1 \qquad (1.17)$$

The resulting iterative method is similar to the one described by McBride. All these optimum least-squares methods will be discussed more fully in chapter III.

I.3 Brief Discussion of Previous Methods

Prony's method and the method of Padé approximants have two things in common. First, each requires the solving of a system of linear simultaneous equations. Second, to find the approximate $\{s_k\}$, one must evaluate the roots of an n^{th} degree polynomial. Each method uses the application of these two operations only once. Hence, each is useful in that it provides a rapid way of obtaining an approximation to a desired waveform. To improve these initial approximations or to make them optimal, either of the linear iterative schemes may be used. These linear iterative methods also require solving a system of linear simultaneous equations and finding the roots of an n^{th} degree polynomial for each iteration. It will be shown in the thesis that the method of McDonough is the better way of improving the approximation.

Regardless of the initial starting point in the approximation, Mode-1, Mode-2, and McDonough's method all converge to the optimal solution in one step if the function f(t) is exactly composed of n exponentials. This suggests that any of the linear methods will converge rapidly to the optimum when the signal is "nearly exponential".

I.4 Plan of the Thesis

It is well known how to find the amplitude coefficients for a least-square approximation to a function on a fixed or known basis. However, computational difficulties arise when the basis elements are highly correlated. In chapter II a closed-form expression is developed for the inverse of some Gram matrices that occur in least-square theory. These new expressions can help to reduce roundoff errors in computing the amplitude coefficients. In particular, the exponential basis is studied. An explicit inverse for the generalized Hilbert matrix, the Gram matrix for an exponential basis, was published in a French Journal in 1960 [14]. Although this result is quite useful in least-square representation by exponentials and polynomials, it appears to have remained unknown to the English literature. Its use is fully explored in regard to finding the amplitudes as well as the $\{s_k\}$.

The successful methods of McDonough and Huggins and of McBride, Schaefgen, and Steiglitz are discussed in detail in chapter III. A new scheme is developed which for the first time enables one to make a meaningful comparison of the methods. In chapter IV, several numerical examples compare the convergence properties of the linear iterative methods. Finally, in chapter V, the new method is extended to deal with imperfectly known signals or sampled data.

II. DETERMINATION OF THE AMPLITUDE COEFFICIENTS IN LEAST-SQUARE APPROXIMATION OF FUNCTIONS BY EXPONENTIALS AND OTHER COMMONLY USED BASIS FUNCTIONS

Suppose $x_1(t)$, $x_2(t)$, ..., $x_n(t)$ denote a finite sequence of conveniently chosen functions defined over some continuous interval (a,b) of t. Let $f_a(t) = \sum_{k=1}^n \alpha_k x_k(t)$ be an approximation to the function f(t). One problem is to find the α_k such that $\int_a^b |f(t) - f_a(t)|^2 \, dt = J$ is a minimum. The standard least-squares procedure yields the following equations:

$$\frac{\partial J}{\partial \alpha_i} = 0 \text{ or } \sum_{j=1}^{n} g_{ij}\alpha_j = f_i \qquad i = 1,2,\ldots,n \tag{2.1}$$

where $g_{ij} = \int_a^b x_i(t)x_j^*(t) \, dt = g_{ji}^*$ are the elements of the Gram matrix $\underline{\underline{G}}$ and $f_i = \int_a^b f(t)x_i^*(t) \, dt$ are the elements of the column $\underline{F}$. Then the best fitting amplitude coefficients are given by the column matrix $\underline{A}$, where

$$\underline{A} = \underline{\underline{G}}^{-1}\underline{F}. \tag{2.2}$$

Theoretically, this algorithm presents no difficulty provided the x_i are linearly independent. If they are not, the matrix $\underline{\underline{G}}$ will be singular but the same minimum error J can be obtained with a set of less than n of the x_i that are independent. When the x_i are highly correlated, but independent (exponentials for example), the matrix $\underline{\underline{G}}$ is "ill-conditioned" and computational difficulties arise in finding the inverse accurately for any sizeable n, as evidenced in [19]-[21]. This difficulty is sometimes reduced by introducing a new basis of orthonormal functions, which are linear combinations of the original basis functions $x_i(t)$ and span the same space. However, these orthonormal functions no longer posess the simple properties of the original basis so this approach is not a cure-all. A method for finding $\underline{\underline{G}}^{-1}$ is still needed.

II.1 Inverse of the Gram Matrix

Let $\phi_1(t),\phi_2(t),\ldots,\phi_n(t)$ be a set of orthonormal functions, which may be determined from the $x_i(t)$ by the Gram-Schmidt procedure. That is,

$$\phi_i(t) = \sum_{k=1}^{i} c_{ik}x_k(t) \qquad i = 1,2,\ldots,n \tag{2.3}$$

and c_{ii} cannot be zero if the x_k are linearly independent. Written in matrix notation, $\underline{\phi}(t) = \underline{\underline{C}}\underline{X}(t)$, where $\underline{\underline{C}}$ is a nonsingular n X n lower triangular matrix and

$$\int_a^b \phi_i(t)\phi_j^*(t)\, dt = \delta_{ij} \tag{2.4}$$

Using Dirac notation, let $\underline{X}|$ denote the column of basis elements and $|\tilde{\underline{X}}$ its adjoint. Then

$$\underline{X}|\tilde{\underline{X}} = \underline{\underline{G}} \tag{2.5}$$

From (2.3)

$$\underline{\phi}| = \underline{\underline{C}}\,\underline{X}| \tag{2.6}$$

$$\underline{X}| = \underline{\underline{C}}^{-1}\underline{\phi}| \tag{2.7}$$

$$|\tilde{\underline{X}} = |\tilde{\underline{\phi}}\,\tilde{\underline{\underline{C}}}^{-1} \tag{2.8}$$

Note that $\tilde{\underline{\underline{C}}}$ is the adjoint of $\underline{\underline{C}}$ which means $\tilde{c}_{ij} = c_{ji}^*$. If $\underline{\underline{C}}$ is real, then $\tilde{\underline{\underline{C}}}$ is the transpose of $\underline{\underline{C}}$.

$$\underline{X}|\underline{\tilde{X}} = \underline{\underline{C}}^{-1}\ \underline{\Phi}|\underline{\tilde{\Phi}}\ \underline{\underline{\tilde{C}}}^{-1}, \qquad (2.9)$$

but

$$\underline{\Phi}|\underline{\tilde{\Phi}} = \underline{\underline{I}} \text{ (identity matrix).}$$

Hence

$$\underline{\underline{G}} = \underline{\underline{C}}^{-1}\underline{\underline{\tilde{C}}}^{-1} \qquad (2.10)$$

or the final result

$$\underline{\underline{G}}^{-1} = \underline{\underline{\tilde{C}}}\ \underline{\underline{C}}. \qquad (2.11)$$

Equation (2.11) is a useful result in two ways. First, it may be used to find explicit expressions for the inverses of some Gram matrices that are ill-conditioned. Second, it can sometimes be used to construct an orthonormal basis by simple inspection. The second application is not one of the main goals of the thesis and therefore, is discussed in Appendix A. The first application will now be demonstrated by finding the inverse of the Hilbert matrix.

II.2 The Generalized Hilbert Matrix

The generalized Hilbert matrix is the n X n Hermitian matrix with elements $h_{ij} = -(s_i+s_j^*)^{-1}$ or

$$\underline{\underline{H}} = -\begin{bmatrix} 1/(s_1+s_1^*) & 1/(s_1+s_2^*) & \cdots & 1/(s_1+s_n^*) \\ 1/(s_2+s_1^*) & 1/(s_2+s_2^*) & \cdots & 1/(s_2+s_n^*) \\ \vdots & \vdots & & \vdots \\ 1/(s_n+s_1^*) & 1/(s_n+s_2^*) & \cdots & 1/(s_n+s_n^*) \end{bmatrix} \qquad (2.12)$$

where the s_i are n complex scalars and $s_i \neq s_j$ if $i \neq j$, and $s_i \neq 0$.

The Hilbert matrix is discussed extensively in the literature [21]-[24]. The inverse of this matrix is shown to have for its elements

$$h^{-1}_{ij} = -\frac{(s_i+s_i^*)(s_j+s_j^*)}{(s_i^*+s_j)}\left\{\prod_{\substack{k=1\\k\neq i}}^{n}\frac{(s_k+s_i^*)}{(s_k^*-s_i^*)}\right\}\left\{\prod_{\substack{k=1\\k\neq j}}^{n}\frac{(s_k^*+s_j)}{(s_k-s_j)}\right\} \qquad (2.13)$$

Proof: Let $x_i(t) = \exp(+s_i t)$, $\mathrm{Re}(s_i < 0)$, then $h_{ij} = \int_0^\infty x_i(t)\, x_j^*(t)\, dt$. As in (2.3), let the orthonormal functions be given by $\underline{\phi}(t) = \underline{\underline{C}}\, \underline{X}(t)$, where

$$c_{ij} = \frac{(-1)^{i+1}(-s_i-s_i^*)^{1/2}(s_j+s_j^*)}{(s_i^*+s_j)}\left\{\prod_{\substack{k=1\\k\neq j}}^{i}\frac{(s_k^*+s_j)}{(s_k-s_j)}\right\} \qquad (2.14)$$

The $\phi_i(t)$ are known to be orthonormal from Kautz's method [13], [18]. The Laplace transform of $\phi_i(t)$ is

$$\phi_i(s) = \frac{(-s_i-s_i^*)^{1/2}}{(s+s_i^*)}\left\{\prod_{k=1}^{i}\frac{(s+s_k^*)}{(s-s_k)}\right\} \qquad (2.15)$$

The c_{ij} in (2.14) correspond to the residues of this transform. From (2.11)

$$h^{-1}_{ij} = \sum_{\max(i,j)}^{n} c^*_{ki}c_{kj} \qquad (2.16)$$

since $c_{im} = 0$ if $i < m$. From (2.14) and (2.16)

$$\begin{aligned} h^{-1}_{in} &= c^*_{ni}c_{nn} \\ &= -\frac{(s_i+s_i^*)(s_n+s_n^*)}{(s_i^*+s_n)}\left\{\prod_{\substack{k=1\\k\neq i}}^{n}\frac{(s_k+s_i^*)}{(s_k^*-s_i^*)}\right\}\left\{\prod_{\substack{k=1\\k\neq n}}^{n}\frac{(s_k^*+s_n)}{(s_k-s_n)}\right\} \end{aligned} \qquad (2.17)$$

Because of the symmetry in the original matrix, if c_{nn} is replaced by c_{nj} in (2.17), the formula must hold for the general term h^{-1}_{ij} (since the order of $s_1,s_2,\ldots,s_n$ in $\underline{\underline{H}}$ can be changed without affecting the form of the equations) and (2.13) is proved. In the special case that all the s_i are real, the formula reduces to

$$h_{ij}^{-1} = - \frac{4s_i s_j}{(s_i+s_j)} \left\{ \prod_{\substack{k=1 \\ k\neq i}}^{n} \frac{(s_k+s_i)}{(s_k-s_i)} \right\} \left\{ \prod_{\substack{k=1 \\ k\neq j}}^{n} \frac{(s_k+s_j)}{(s_k-s_j)} \right\} \tag{2.18}$$

This result agrees with Gastinel [14] who found this expression for the inverse of a generalized Hilbert matrix by a rather tedious application of Lagrange's interpolation polynomial. Appendix B gives another interesting explicit inverse using Laguerre functions.

II.3 <u>Roundoff Errors in the Amplitude Coefficients Using a Fixed Exponential Basis</u>

Let f(t) be a piecewise continuous real function having finite energy in the semi-infinite interval, i.e.

$$\int_0^\infty f^2(t)\ dt < \infty. \tag{2.19}$$

We wish to find the amplitude coefficients $\{\alpha_k\}$ that will minimize the mean-square error,

$$J = \int_0^\infty [f(t) - \sum_{k=1}^{n} \alpha_k \exp(s_k t)]^2\ dt \tag{2.20}$$

for a specified set of exponential functions, having $\{s_k\}$ with negative real parts. From (2.1), the simultaneous equations for determining the $\{\alpha_k\}$ are†

† Since f(t) is real, the s_k must occur in complex conjugate pairs. Hence, there is no loss in generality if every s_k^* is replaced by s_k in (2.21) and to simplify the typography this will be done throughout the remainder of the thesis.

$$\begin{bmatrix} F(-s_1^*) \\ F(-s_2^*) \\ \vdots \\ F(-s_n^*) \end{bmatrix} = - \begin{bmatrix} 1/(s_1+s_1^*) & 1/(s_1+s_2^*) & \ldots & 1/(s_1+s_n^*) \\ 1/(s_2+s_1^*) & 1/(s_2+s_2^*) & \ldots & 1/(s_2+s_n^*) \\ \vdots & \vdots & & \vdots \\ 1/(s_n+s_1^*) & 1/(s_n+s_2^*) & \ldots & 1/(s_n+s_n^*) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \qquad (2.21)$$

where F(s) is the Laplace transform of f(t). In matrix form these equations are $\underline{F} = \underline{\underline{H}}\ \underline{A}$, where $\underline{\underline{H}}$ is the generalized Hilbert matrix, and their solution is $\underline{A} = \underline{\underline{H}}^{-1}\ \underline{F}$. However, the Hilbert matrix is notoriously ill-conditioned and computation of $\underline{\underline{H}}^{-1}$ by any of the standard methods (Gauss-Jordan, Seidel's method, method of Crout, etc.) encounters serious roundoff difficulties for n greater than 5 or so, even when double-precision arithmetical operations are used.

The rapid growth of roundoff errors with increasing n may be demonstrated by comparing the inverse of $\underline{\underline{H}}$ (for $s_i=i \quad i=1,2,\ldots,n$ with n=5 and 7) calculated by the explicit formula (2.13) with the inverse obtained by the method of Crout [15]. All calculations were made by an IBM 7094 having approximately 8 significant decimal digits. Table 2.1 shows that only 3-place accuracy is attained in many of the elements of the inverse matrix for n=5, and a complete loss of significant results for n=7 using Crout's method. For still larger n the results are meaningless. On the other hand, the explicit formula achieves 7-place accuracy (for both n=5 and 7). (This was validated by double-precision calculations.) Since a detailed analysis of roundoff errors arising in inversion of matrices on computers is given in references [16] and [17], this topic will not be discussed further here. Moreover, the explicit formula for inverting the Hilbert matrix has so reduced these errors in finding the inverse as to make them insignificant for modest n.

Table 2.1

Inverse of the Hilbert Matrix $h_{ij}=1/(i+j)$ for n=5 and 7 by Method of Crout and by Explicit Formula on Computer Having 8 Significant Figure Accuracy.

$$\begin{bmatrix} 450.0 & -4200.0 & 12600.0 & -15120.0 & 6300.0 \\ -4200.0 & 44100.0 & -141120.0 & 176400.0 & -75600.0 \\ 12600.0 & -141120.0 & 470400.0 & -604800.0 & 264600.0 \\ -15120.0 & 176400.0 & -604800.0 & 793800.0 & -352800.0 \\ 6300.0 & -75600.0 & 264600.0 & -352800.0 & 158760.0 \end{bmatrix}$$

Explicit Formula n=5 (Exact Inverse)

$$\begin{bmatrix} 449.5 & -4194.8 & 12582.6 & -15097.7 & 6290.2 \\ -4194.7 & 44041.4 & -140925.5 & 176150.4 & -75490.9 \\ 12582.4 & -140924.6 & 469751.6 & -603968.1 & 264236.6 \\ -15097.4 & 176148.4 & -603965.6 & 792730.3 & -352332.8 \\ 6290.1 & -75489.8 & 264234.7 & -352331.8 & 158555.5 \end{bmatrix}$$

Method of Crout n=5 (3 Significant Places)

$$\begin{bmatrix} 1568.0 & -28224.0 & 176400.0 & -517440.0 & 776160.0 & -576576.0 & 168168.0 \\ -28224.0 & 571536.0 & -3810239.9 & 11642399.4 & -17962559.5 & 13621407.4 & -4036031.8 \\ 176400.0 & -3810239.9 & 26460000.0 & -83159995.0 & 130977000.0 & -100900795.0 & 30270238.3 \\ -517440.0 & 11642399.4 & -83159994.0 & 266804976.0 & -426887972.0 & 332972608.0 & -100900791.0 \\ 776160.0 & -17962559.8 & 130977000.0 & -426887976.0 & 691558560.0 & -544864296.0 & 166486310.0 \\ -576576.0 & 13621407.4 & -100900795.0 & 332972612.0 & -544864288.0 & 432864400.0 & -133189043.0 \\ 168168.0 & -4036031.8 & 30270238.0 & -100900791.0 & 166486310.0 & -133189045.0 & 41225180.5 \end{bmatrix}$$

Explicit Formula n=7 (7 Significant Places)

$$\begin{bmatrix} 4761.1 & -103971.1 & 726586.2 & -2314999.1 & 3707186.0 & -2892320.9 & 877794.2 \\ -103561.5 & 2137593.6 & -16787078.8 & 54154259.5 & -87083785.0 & 68230085.0 & -20757522.0 \\ 732761.8 & -16850767.1 & 122272628.0 & -397021548.0 & 641270576.0 & -504038756.0 & 153710414.0 \\ -2367352.0 & 54529637.5 & -398246288.0 & 1298915664.0 & -2104897104.0 & 1658586496.0 & -506791418.0 \\ 3761109.5 & -87915085.0 & 644884624.0 & -2110217176.0 & 3428246104.0 & -2706784864.0 & 828439437.0 \\ -2961105.2 & 69031672.0 & -507957548.0 & 1666268928.5 & -2712448000.0 & 2145180288.0 & -657672514.0 \\ 833829.6 & -21040351.5 & 155184792.0 & -510044460.0 & 831636392.0 & -658626632.0 & 202104144.0 \end{bmatrix}$$

Method of Crout n=7 (No Significance)

****

Another source of roundoff error occurs in the computation of the $\{\alpha_k\}$ when $\underline{\underline{H}}^{-1}$ is multiplied by $\underline{F}$. From (2.21) and the explicit formula for $\underline{\underline{H}}^{-1}$, the k^{th} amplitude coefficient can be expressed as

$$\alpha_k = - 4s_k T_k^n \left[\sum_{i=1}^{n} \frac{s_i T_i^n}{s_i + s_k} F(-s_i) \right] \tag{2.22}$$

where

$$T_k^n = \prod_{\substack{m=1 \\ m \neq k}}^{n} \frac{(s_m + s_k)}{(s_m - s_k)} \tag{2.23}$$

The estimation of a roundoff error in evaluating equation (2.22) may be illustrated by considering the approximation of a square pulse, $f(t) = u_{-1}(t) - u_{-1}(t-1)$, where $u_{-1}(t)$ is the unit step,

$$u_{-1}(t) = 1 \quad t \geq 0$$
$$= 0 \quad t < 0$$

and thus

$$F(s) = (1-e^{-s})/s. \tag{2.24}$$

Again let $s_i = i$ $i=1,2,\ldots,n$. Columns 1, 2, and 6 of Table 2.2 summarize the results using equation (2.22) with $n=5,7,$and 9. (The error is estimated by comparing these results with those obtained by double-precision calculations.) Notice that α_k and T_k^n exhibit nearly the same order of magnitude for almost all n and k. This implies that the sum of the n terms within the brackets of (2.22) must be roughly of unit magnitude. Also, all of the $F(-s_k)$ are less than 1. (In chapter V it is shown that for any normalized $f(t)$, $|F(-s_k)| \leq [2\mathrm{Re}\{-s_k\}]^{-1/2}$.) Therefore, the number of significant decimal digits lost in each of the α_k computed by this method, which forms small differences of very large numbers, may be expected to be the same as the number of places to the left of the decimal point in the largest T_k^n. In the example provided by Table 2.2, for $n=5$,

T_4^5 has the largest magnitude of 315. Thus, a loss of about three significant places may be expected; for n=9, T_6^9 = -210210 indicating a loss of six significant places in each α_k. These predictions agree with the actual accuracies obtained in Table 2.2.

Table 2.2

Amplitude Coefficients of exp(-kt) k=1,2,...,n for a Least-Square Approximation of the Square Pulse

α_k From Eq.(2.22)	Error	α_k From Eq.(2.32)	Error	α_k Double-Precision	T_k^n
0.296	0.000	0.296	-0.000	0.296	15.000
-12.909	-0.001	-12.907	0.001	-12.908	-105.000
80.120	0.004	80.115	-0.002	80.117	280.000
-126.475	-0.004	-126.468	0.003	-126.471	-315.000
60.312	0.002	60.309	-0.001	60.310	126.000

n=5 Loss of 3-Places

1.204	0.010	1.191	-0.004	1.194	28.000
-19.080	-0.170	-18.825	0.085	-18.910	-378.000
35.712	1.308	33.765	-0.640	34.404	2100.000
263.830	-4.017	269.987	2.141	267.846	-5775.000
-909.182	7.754	-920.482	-3.547	-916.935	8316.000
993.607	-5.937	1002.392	2.848	999.544	-6006.000
-364.836	1.576	-367.297	-0.885	-366.412	1716.000

n=7 Loss of 5-Places

-2.543	0.146	-2.601	0.088	-2.689	45.001
103.931	-1.969	102.238	-3.662	105.900	-990.000
-1211.336	35.434	-1198.521	48.309	-1246.839	9240.001
6037.525	-315.298	6056.937	-295.886	6352.826	-45044.998
-15819.637	466.796	-15309.476	976.957	-16286.433	126126.000
20861.905	-1776.682	20792.096	-1846.490	22638.586	-210209.992
-16344.003	660.031	-15005.783	1998.252	-17004.036	205919.992
5501.799	-731.881	5082.465	-1151.214	6233.680	-109395.995
-680.740	109.115	-516.197	273.658	-789.855	24309.999

n=9 Loss of 7-Places

●●●●●●●●

A procedure that is often used to avoid the inversion of the ill-conditioned matrix of equation (2.21) is achieved by introducing orthogonalized combinations of the original exponentials. The orthogonalizing procedure may be implemented in practical filters by a method due to Kautz [13], [18] which is based on the traditional Gram-Schmidt procedure applied to exponential functions. Rewriting (2.14), the orthonormal functions are

$$\phi_i(t) = \sum_{k=1}^{i} c_{ik} \exp(s_k t) \tag{2.25}$$

where

$$c_{ik} = (-1)^{i+1} \frac{(s_k+s_k^*)\sqrt{-(s_i+s_i^*)}}{s_i+s_k} T_k^i \qquad i \geq k \tag{2.26}$$

Then

$$f_a(t) = \sum_{k=1}^{n} d_k \, \phi_k(t). \tag{2.27}$$

As is well-known from the theory of orthonormal basis, the expansion coefficients

$$d_k = \int_0^\infty f(t) \, \phi_k(t) \, dt \tag{2.28}$$

automatically yield a least-squares fit. Equation (2.28) may be expressed in matrix form as

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} = \begin{bmatrix} c_1 & 0 & 0 \ldots & 0 \\ c_{21} & c_{22} & 0 \ldots & 0 \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & c_{n3} \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} F(-s_1) \\ F(-s_2) \\ \vdots \\ F(-s_n) \end{bmatrix} \tag{2.29}$$

or $\underline{D} = \underline{\underline{C}}\,\underline{F}$. When the signal coordinates on the orthonormal basis are transformed to find the coordinates on the original exponential basis one gets

$$\begin{array}{llll} \alpha_1 = c_{11}d_1 & + c_{21}d_1 & + c_{31}d_1 & + \cdots \\ \alpha_2 = & c_{22}d_2 & + c_{32}d_3 & + \cdots \\ \alpha_3 = & & c_{33}d_3 & + \cdots \\ \cdots & \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots \end{array} \qquad (2.30)$$

Equation (2.30) is easily verified using (2.2), (2.11) and (2.29). Hence, explicit equations for the $\{\alpha_k\}$ can be obtained in two simple steps by combining[†] (2.29) and (2.30),

$$\underline{A} = \underline{\underline{\tilde{C}}}\,(\underline{\underline{C}}\,\underline{F}) \qquad (2.31)$$

or

$$\alpha_k = \sum_{i=k}^{n} c_{ik} \left[\sum_{j=1}^{i} c_{ij}\, F(-s_j) \right] \qquad (2.32)$$

To minimize the number of arithmetical operations required in evaluating $\underline{A} = \underline{\underline{\tilde{C}}}\,\underline{\underline{C}}\,\underline{F}$, the product $\underline{\underline{C}}\,\underline{F}$ should be formed first. This requires n(n+1) multiplications, whereas if $\underline{\underline{\tilde{C}}}\,\underline{\underline{C}}$ is formed first, roughly $n^3/2$ multiplications are needed to find all the $\{\alpha_k\}$.

Although $\underline{A}$ may be evaluated by either equation (2.22) or (2.32) (which are theoretically equivalent), equation (2.22) is computationally preferred for three reasons. First, it is much simpler, requiring only a single summation rather than the double summation

[†] Of course, if the $\underline{\underline{\tilde{C}}}\,\underline{\underline{C}}$ in (2.31) is combined and simplified, one obtains the explicit expression for the inverse of the generalized Hilbert matrix obtained earlier. Apparently, this way of finding the inverse of a Gram matrix has not appeared previously in the literature.

of (2.32). Second, only n T_k^n are needed in the first method, whereas in the second, n(n+1)/2 different c_{ij} must be evaluated.[†] Third, the method of equation (2.22) provides a simple estimate of the number of significant places that will be lost due to roundoff even before the actual computation is made.

By observing the magnitudes of the T_k^n, in Table 2.2, we have already noted that the percent roundoff error corresponds to the magnitude of the largest T_k^n. Table 2.2 shows that the accuracy of either method is about the same, so the choice rests entirely on which offers the greatest computational advantage: this is the method of equation (2.22).

Thus far, we have presented two methods for determining the amplitude coefficients. For single-precision computation sizeable numerical errors arise in both methods for n greater than 4,

[†] Some simplification is possible in evaluating these c_{ij} by making use of a recursion relation which requires the calculation of only the n c_{kk}, all other quantities being obtained from these. The relation obtained from (2.23) and (2.26) is

$$c_{i+1,k} = -\sqrt{\frac{(s_{i+1}+s_{i+1}^*)}{(s_i+s_i^*)}}\ \frac{s_i+s_k}{s_{i+1}-s_k}\ c_{ik} \qquad i \geq k$$

Even using this recursion, the computation of the c_{nk} alone requires at least as much work as the T_k^n.

and for n greater than 15 (maximum $T_k^n > 10^{10}$ with $s_i = i$) even double-precision arithmetic may not be adequate. There remains a real need for further detailed study of the computational aspects of these methods.

II.4 The Vandermonde Matrix

The Vandermonde matrix arises in many branches of applied mathematics. In control theory one encounters the equation $\dot{\underline{X}}(t) = \underline{\underline{A}}\,\underline{X}(t) + \underline{D}\,m(t)$ [27], which may be simplified by transforming the state vector $\underline{X}$ to $\underline{Y} = \underline{\underline{V}}^{-1}\,\underline{X}$ where $\underline{\underline{V}}$ is the Vandermonde matrix.

In numerical interpolation by polynomials of a function defined by a set of n ordered pairs of real or complex numbers (s_k, z_k) with all the s_k distinct, one seeks a unique polynomial

$$N(s) = a_1 + a_2\,s + \dots + a_n\,s^{n-1} \qquad (2.33)$$

for which

$$N(s_k) = z_k \qquad k = 1, 2, \dots, n. \qquad (2.34)$$

The conditions (2.34) form a system of n linear equations in the a_i coefficients of the polynomial,

$$\begin{bmatrix} 1 & s_1 & s_1^2 & \dots & s_1^{n-1} \\ 1 & s_2 & s_2^2 & \dots & s_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & s_n & s_n^2 & \dots & s_n^{n-1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}. \qquad (2.35)$$

The matrix of this system is named after Vandermonde and is shown to be non-singular provided all the s_k are distinct [28].

The Vandermonde matrix also arises in least-square approximation using exponentials over the semi-infinite interval as we now show. By solving equation (2.21), one obtains the best fitting approximation

$$F_a(s) = \frac{\alpha_1}{s-s_1} + \frac{\alpha_2}{s-s_2} + \dots + \frac{\alpha_n}{s-s_n} \tag{2.36}$$

which has the properties ennumerated by Aigrain and Williams [12], that

$$F(-s_k) = F_a(-s_k) \qquad k = 1,2,\dots,n. \tag{2.37}$$

Equation (2.36) may also be written as the rational fraction

$$F_a(s) = \frac{a_1+a_2s+\dots+a_ns^{n-1}}{b_1+b_2s+\dots+b_ns^{n-1}+s^n} \tag{2.38}$$

$$= \frac{N(s)}{(s-s_1)(s-s_2)\dots(s-s_n)} = \frac{N(s)}{D(s)} \,. \tag{2.39}$$

When the $\{a_k\}$ are optimally chosen, equation (2.37) is satisfied. Then,

$$D(-s_k)\;F(-s_k) = D(-s_k)\;F_a(-s_k) = N(-s_k) \quad k=1,2,\dots,n. \tag{2.40}$$

In matrix form,

$$\begin{bmatrix} D(-s_1)\;F(-s_1) \\ D(-s_2)\;F(-s_2) \\ \vdots \\ D(-s_n)\;F(-s_n) \end{bmatrix} = \begin{bmatrix} 1 & (-s_1) & (-s_1)^2 \dots & (-s_1)^{n-1} \\ 1 & (-s_2) & (-s_2)^2 \dots & (-s_2)^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (-s_n) & (-s_n)^2 \dots & (-s_n)^{n-1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \tag{2.41}$$

Equation (2.41) also exhibits the Vandermonde matrix $\underline{\underline{V}}$ with elements

$$v_{ij} = (-s_i)^{j-1} \quad i,j=1,2,\dots,n. \tag{2.42}$$

To solve (2.41) for the coefficients of the numerator polynomial, a closed-form expression for the inverse $\underline{\underline{V}}^{-1}$ of the Vandermonde matrix is needed. An explicit expression for the inverse of this important matrix is given by Tou [29] and the result may be summarized by the following theorem:

Theorem 1 - Let $\underline{\underline{V}}$ be the Vandermonde matrix with elements $v_{ij}=(-s_i)^{j-1}$ and $\underline{\underline{V}}^{-1}$ be its inverse with elements v_{ij}^{-1}. Then the generating function for these inverse elements is the Lagrange interpolating polynomial [30],

$$L_j(s) = \prod_{\substack{k=1\\k\neq j}}^{n} \frac{(s+s_k)}{(s_k-s_j)} = \sum_{i=1}^{n} v_{ij}^{-1} s^{i-1} \qquad s_k\neq s_j \text{ if } k\neq j.$$

for which[†]

$$L_j(-s_k) = \delta_{jk}$$

Theorem 1 will assist us, in chapter III, in making a direct comparison of two recently developed methods for exponential approximation.

[†] Usually the Lagrange interpolating polynomial is written

$$L_j(s) = \prod_{\substack{k=1\\k\neq j}}^{n} \frac{(s-s_k)}{(s_j-s_k)}$$. This difference is due to the negative elements

$\{-s_i\}$ in $\underline{\underline{V}}$.

III. DETERMINATION OF THE MATCHED EXPONENTS OF A DEFINED ANALYTIC FUNCTION

In the previous chapter, two methods were examined to find the amplitude coefficients in a least-square approximation of a function f(t) by a specified set of exponential functions. These linear least-squares procedures can always be carried out given sufficient time and precision to determine accurately the amplitude coefficients. In contrast, finding the complex frequencies of the set of exponential components to best match the specified function f(t) is much more difficult. Our attack on finding this set of matched exponents begins with the equations of Aigrain and Williams [12].

III.1 The Equations of Aigrain and Williams

For a given f(t), $t \geq 0$, the necessary conditions on the 2n parameters $\{\alpha_k, s_k\}$ to minimize the functional

$$J = \int_0^\infty [f(t) - \sum_{k=1}^n \alpha_k \exp(s_k t)]^2 \, dt \tag{3.1}$$

are expressed by the two sets of n equations

$$\frac{\partial J}{\partial \alpha_j} = \int_0^\infty 2\,[f(t) - \sum_{k=1}^n \alpha_k \exp(s_k t)]\,[-\exp(s_j t)]\,dt = 0,$$

or

$$\sum_{k=1}^n \alpha_k \int_0^\infty \exp((s_j+s_k)t)\,dt = \int_0^\infty f(t)\exp(s_j t)\,dt \tag{3.2a}$$

$$j=1,2,\ldots,n.$$

and

$$\frac{\partial J}{\partial s_j} = \int_0^\infty 2[f(t) - \sum_{k=1}^n \alpha_k \exp(s_k t)]\,[-\alpha_j t \exp(s_k t)]\,dt = 0,$$

or

$$-\sum_{k=1}^{n} \alpha_k \int_0^\infty t \exp((s_j+s_k)t)\, dt = -\int_0^\infty f(t)\, t \exp(s_j t)\, dt \qquad (3.2b)$$

$$j=1,2,\ldots,n.$$

These conditions for stationarity of the integrated squared error may be written in the frequency domain as

$$\left.\begin{aligned} -\sum_{k=1}^{n} \frac{\alpha_k}{s_j+s_k} &= F(-s_j) \\ -\sum_{k=1}^{n} \frac{\alpha_k}{(s_j+s_k)^2} &= F'(-s_j) \end{aligned}\right\} \quad j=1,2,\ldots,n \qquad \begin{aligned} &(3.3a) \\ &(3.3b) \end{aligned}$$

or even more simply as

$$\left.\begin{aligned} F_a(-s_j) &= F(-s_j) \\ F_a'(-s_j) &= F'(-s_j) \end{aligned}\right\} \quad j=1,2,\ldots,n \qquad \begin{aligned} &(3.4a) \\ &(3.4b) \end{aligned}$$

where as usual[†],

$$F(s) = \int_0^\infty f(t) \exp(-st)\, dt \qquad (\mathrm{Re}\{s\}>\sigma_o) \qquad (3.5a)$$

$$F'(s) = \frac{d}{ds}\,[F(s)] \qquad (3.5b)$$

and

$$F_a(s) = \sum_{k=1}^{n} \frac{\alpha_k}{s-s_k} \qquad (3.5c)$$

(Equations (3.4) reveal that in approximation theory, the important information of the signal is contained at the mirror images of the poles of $F_a(s)$ which are all points such that Re{s}>0. This suggests that the most useful information about F(s) and $F_a(s)$ is in the right

[†] $\sigma_o = 0$ if $\int_0^\infty f^2(t)\, dt < \infty$.

half plane and not at the poles of $F_a(s)$! To further demonstrate this point, consider the following two functions,

$$g_1(t) = \exp(-\alpha t)\ u_{-1}(t) \qquad \alpha > 0$$

$$g_2(t) = \exp\ (-\alpha t)\ [u_{-1}(t) - u_{-1}(t-\tau)] \qquad \tau > 0$$

Then the corresponding Laplace transforms are

$$G_1(s) = \frac{1}{s+\alpha}$$

$$G_2(s) = \frac{1-\exp(-(s+\alpha)\tau)}{s+\alpha} .$$

Notice that $G_2(s)$ does not have any poles even for arbitrarily large τ. This means that the pole of $G_1(s)$ in the left half plane is due solely to the tail end of the exponential which is a negligible part of the function $g_1(t)$ for large $\alpha\tau$.)

The 2n equations (3.4) were formulated by Aigrain and Williams in 1948. Despite their simple appearance, closed-form solution of these nonlinear equations is impossible except in trivial cases. Two ways that have been used to solve these equations are gradient methods and "linear iterative schemes". These methods will now be discussed.

III.1.1 Gradient Methods

One straightforward way of finding the matched exponents is based on the method of steepest descent. That is, one finds a suitable scalar function of the 2n parameters $\{\alpha_k, s_k\}$ which has a relative minimum for values of these parameters that satisfy the Aigrain and Williams equations. Clearly, a suitable function is the integrated squared-error J defined in (3.1). The gradient of J is computed at some initial point in parameter space and then the parameter point is perturbed in the direction of the negative gradient. The process is repeated until the gradient is approximately zero.

Let $f(t)$ be normalized so that

$$||f(t)||^2 = \langle f(t),\tilde{f}(t)\rangle = \int_0^\infty f(t)f^*(t)\,dt = 1 \qquad (3.6)$$

Then

$$J = ||f-f_a||^2 = 1-2\sum_{k=1}^{n} \alpha_k F(-s_k) + \sum_{k=1}^{n} \alpha_k F_a(-s_k). \qquad (3.7)$$

The dependence of J on the $\{\alpha_k\}$ may be suppressed by using (3.4a) and its equivalent form (2.21). Under the constraint of equation† (3.4a),

† Equation (3.8) has an interesting geometric interpretation. By definition, the α_k are the coordinates of f_a on the oblique basis $\underline{B}|$ whose elements are $\{\exp(s_k t)\}$. The reciprocal basis is defined as $\tilde{\underline{D}}|$ where $|\underline{D} = |\tilde{\underline{B}}\,(\underline{B}|\tilde{\underline{B}})^{-1}$. A "vector" $\langle f_a|$ may be also written as a linear combination of the dual basis elements,

$$\langle f_a| = \sum_{k=1}^{n} g_k \langle \tilde{D}_k| = \langle \underline{G}\,\tilde{\underline{D}}|$$

The square of the length of $\langle F_a|$ is

$$||f_a||^2 = \langle f_a,\tilde{f}_a\rangle = \langle \underline{A}\;\underline{B} \mid \underline{D}\;\tilde{\underline{G}}\rangle = \langle \underline{A},\tilde{\underline{G}}\rangle$$

$$= \sum_{k=1}^{n} \alpha_k\, g_k.$$

However, it is a well-known fact that $J = ||e||^2 = ||f||^2 - ||f_a||^2$ in a least-square approximation. Hence, the $\{F(-s_k)\}$ are the coordinates of $f(t)$ on the reciprocal basis of $\underline{B}|$ i.e.

$$\langle f_a| = \sum_{k=1}^{n} F(-s_k) \langle \tilde{D}_k|.$$

$$J = 1 - \sum_{k=1}^{n} \alpha_k \, F(-s_k) \tag{3.8}$$

and from (2.21)

$$J = 1 - \sum_{k=1}^{n} \sum_{j=1}^{n} F(-s_k) \, h_{kj}^{-1} \, F(-s_j) \tag{3.9}$$

where h_{kj}^{-1} are the elements of the inverse of the Hilbert matrix defined in (2.13). In matrix form

$$\begin{aligned} J &= 1 - \tilde{\underline{F}} \, \underline{\underline{H}}^{-1} \, \underline{F} = 1 - \tilde{\underline{F}} \, \tilde{\underline{\underline{C}}} \, \underline{\underline{C}} \, \underline{F} \\ &= 1 - (\widetilde{\underline{\underline{C}} \, \underline{F}}) \, \underline{\underline{C}} \, \underline{F} \end{aligned} \tag{3.10}$$

where the c_{ik} are defined in (2.14). Hence, equation (3.10) gives an explicit expression for the integrated squared error in terms of the $\{s_k\}$ only.

For the scalar function J, however, gradient methods have two serious shortcomings. First, this error is insensitive to changes in the $\{s_k\}$ and as a result convergence to the minimum is slow. Second, because of the correlation between exponentials the error can be reduced to near its minimum value for a wide range of $\{s_k\}$. In several cases tried, descent methods converged to values other than the minimum. (Box [31] has shown with several examples why gradient methods don't always converge to the minimum.) As n increases, convergence to the matched exponents by gradient methods becomes difficult to attain since the measure of dependence between the set of exponentials increases so rapidly with n.

A better method of attack, achieved by working directly with the Aigrain-Williams equations, will now be considered.

III.1.2 Direct Method of Elimination of the $\{\alpha_i\}$ From the Equations of Aigrain and Williams

In matrix form equations (3.3) are:

$$\underline{F} = \underline{\underline{H}}\,\underline{A} \qquad (3.11a)$$

and

$$\underline{F}' = -\underline{\underline{G}}\,\underline{A} \qquad (3.11b)$$

with

$$\underline{F} = \begin{bmatrix} F(-s_1) \\ F(-s_2) \\ \vdots \\ F(-s_n) \end{bmatrix}, \quad \underline{A} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \quad \underline{F}' = \begin{bmatrix} F'(-s_1) \\ F'(-s_2) \\ \vdots \\ F'(-s_n) \end{bmatrix}$$

and $\underline{\underline{G}}$ and $\underline{\underline{H}}$ are n X n matrices with elements

$$g_{ij} = \frac{1}{(s_i+s_j)^2}, \quad h_{ij} = \frac{-1}{(s_i+s_j)} \qquad i,j=1,2,\ldots,n.$$

Since the $\{\alpha_k\}$ may be expressed as a function of the $\{s_k\}$ only, (2.21), the 2n equations (3.11) may be reduced to n equations involving the n unknown $\{s_k\}$ alone. By matrix inversion the n new equations become

$$\underline{F}' = -\underline{\underline{G}}\,\underline{\underline{H}}^{-1}\,\underline{F} = \underline{\underline{B}}\,\underline{F} \qquad (3.12)$$

and the $\{\alpha_k\}$ are eliminated. However, equation (3.12) is hopelessly nonlinear in $\{s_k\}$ and in its present form has been found to be worthless for computing these exponents. The next two theorems will help put (3.12) in more useful form.

III.1.2.1 Explicit Inverse of the Hilbert Matrix

In chapter II, a derivation for an explicit expression for the inverse of the generalized Hilbert matrix was derived using orthonormal exponentials and the restriction that the real part of every s_k be positive. This constraint was introduced simply to insure that the exponential components could be normalized. Except for that, it seems to be unnecessary and may be removed by "analytic continuation", to yield the more general theorem:

<u>Theorem 2</u> - Let $h_{ij} = -1/(s_i+s_j)$ be an element of the n X n generalized Hilbert matrix $\underline{\underline{H}}$ associated with the set of n scalars $\{s_k\}$ with $s_i \neq \pm s_j$ for all i and j $(i\neq j)$ and $s_i\neq 0$ $i=1,2,\ldots,n$. Define $\underline{\underline{D}}$ as the n X n matrix with elements

$$d_{ij} = -\frac{4s_i s_j}{s_i+s_j} T_i^n T_j^n$$

where

$$T_m \equiv T_m^n = \prod_{\substack{k=1\\k\neq m}}^{n} \frac{(s_k+s_m)}{(s_k-s_m)} . \qquad (3.13)$$

Then $\underline{\underline{D}} = \underline{\underline{H}}^{-1}$

<u>Proof by induction</u>

Let $\underline{\underline{\Delta}}$ be the product matrix $\underline{\underline{\Delta}} = \underline{\underline{H}}\,\underline{\underline{D}}$. We wish to prove that $\underline{\underline{\Delta}}$ is the unit matrix with elements δ_{ij}. Define

$$\Delta_{ij} \equiv \Delta_{ij}^n = \sum_{k=1}^{n} h_{ik}\, d_{kj}$$

$$= \sum_{k=1}^{n} \frac{4s_k s_j T_k^n\, T_j^n}{(s_i+s_k)(s_k+s_j)} . \qquad (3.14)$$

Then

$$\Delta_{ij}^{n-1} = \sum_{k=1}^{n-1} \frac{4s_k s_j T_k^{n-1}\, T_j^{n-1}}{(s_i+s_k)(s_k+s_j)} .$$

From (3.13) it follows that

$$T_k^n = \frac{(s_n+s_k)}{(s_n-s_k)} T_k^{n-1}$$

and thus

$$\Delta_{ij}^{n-1} = \sum_{k=1}^{n} \frac{4s_k s_j T_k^n T_j^n}{(s_i+s_k)(s_k+s_j)} \frac{(s_n-s_k)}{(s_n+s_k)} \frac{(s_n-s_j)}{(s_n+s_j)} \tag{3.15}$$

so

$$\Delta_{ij}^n - \Delta_{ij}^{n-1} = \sum_{k=1}^{n} \left[\frac{4s_k s_j T_k^n T_j^n}{(s_i+s_k)(s_k+s_j)} \left\{ 1 - \frac{(s_n-s_k)(s_n-s_j)}{(s_n+s_k)(s_n+s_j)} \right\} \right]$$

$$= \frac{2s_n}{(s_n+s_j)} \sum_{k=1}^{n} \frac{4s_k s_j T_k^n T_j^n}{(s_i+s_k)(s_k+s_n)}$$

or

$$\Delta_{ij}^n - \Delta_{ij}^{n-1} = \left[\frac{2s_j T_j^n}{(s_n+s_j) T_n^n} \right] \Delta_{in}^n \qquad i,j=1,2,\ldots,n-1. \tag{3.16}$$

For (3.16) to hold when $j=n$ and $i \neq n$, Δ_{in}^{n-1} must be defined to be zero. For (3.16) to hold when i and j are both equal to n, one must subtract Δ_{nn}^{n-1} from the right hand side of (3.16) whenever $i=n$. With this modification, so that it may be applied generally, equation (3.16) becomes

$$\Delta_{ij}^n - \Delta_{ij}^{n-1} = \frac{2s_j T_j^n}{(s_n+s_j) T_n^n} (\Delta_{in}^n - \delta_{in} \Delta_{nn}^{n-1}) \quad i,j=1,2,\ldots,n. \tag{3.16a}$$

We now assert as the inductive hypothesis

$$\Delta_{ij}^{n-1} = \delta_{ij}, \tag{3.17}$$

which is readily shown to be true for $n=2$ and 3. To establish validity for larger n, first substitute (3.17) in (3.16a) to obtain

$$[\Delta_{ij}^n - \delta_{ij}] - \frac{2s_j T_j^n}{(s_n+s_j) T_n^n} [\Delta_{in}^n - \delta_{in}] = 0. \quad i,j=1,2,\ldots,n \tag{3.18}$$

Then for the n X n matrix $\underline{\underline{\Delta}}$

$$\Delta_{ij} = \delta_{ij} + \frac{2s_jT_j}{(s_n+s_j)T_n}(\Delta_{in} - \delta_{in}) \tag{3.19}$$

From the symmetry in equation (3.14), if the subscripts j and n are interchanged equation (3.19) must hold. Hence,

$$\Delta_{in} = \delta_{in} + \frac{2s_nT_n}{(s_n+s_j)T_j}(\Delta_{ij} - \delta_{ij}) \tag{3.20}$$

Substituting (3.20) in (3.19) one gets

$$\left[1 - \frac{4s_ns_j}{(s_n+s_j)^2}\right][\Delta_{ij} - \delta_{ij}] = 0$$

Thus, it is necessary that

$$\Delta_{ij} = \delta_{ij} \qquad i,j=1,2,\ldots,n$$

and Theorem 2 is proved by induction. Q. E. D.

III.1.2.2 <u>Simplification of $\underline{\underline{B}} = -\underline{\underline{G}}\,\underline{\underline{H}}^{-1}$.</u>

<u>Theorem 3</u> - Let $\underline{\underline{G}}$ and $\underline{\underline{H}}^{-1}$ be n X n symmetric matrices with elements

$$g_{ij} = \frac{1}{(s_i+s_j)^2} \quad \text{and} \quad h^{-1}_{ij} = \frac{-4s_is_j}{(s_i+s_j)} T_i^n\, T_j^n.$$

Then the elements of the matrix $\underline{\underline{B}} = -\underline{\underline{G}}\,\underline{\underline{H}}^{-1}$ of equation (3.15) are

$$B^n_{ij} = -\sum_{k=1}^{n} g_{ik}\, h^{-1}_{kj}$$

$$= \frac{s_j\, T^n_j}{s_i(s_i-s_j)T^n_i} \qquad i\neq j\ , \tag{3.21a}$$

$$B^n_{ii} = -\frac{1}{2s_i} + \sum_{\substack{k=1\\k\neq i}}^{n} \frac{2s_k}{s_i^2-s_k^2}\ . \tag{3.21b}$$

<u>Proof</u>

Formula (3.21) holds for n=2 since $-\underline{\underline{G}}\,\underline{\underline{H}}^{-1} =$

$$\left[\begin{array}{c|c} \frac{1}{(2s_1)^2} & \frac{1}{(s_1+s_2)^2} \\ \hline \frac{1}{(s_2+s_1)^2} & \frac{1}{(2s_2)^2} \end{array}\right] \left[\begin{array}{c|c} 2s_1\left(\frac{s_1+s_2}{s_1-s_2}\right)^2 & -\frac{4s_1s_2}{(s_1+s_2)}\left(\frac{s_1+s_2}{s_1-s_2}\right)^2 \\ \hline -\frac{4s_1s_2}{(s_1+s_2)}\left(\frac{s_1+s_2}{s_1-s_2}\right)^2 & 2s_2\left(\frac{s_1+s_2}{s_1-s_2}\right)^2 \end{array}\right] =$$

$$\left[\begin{array}{c|c} -\frac{1}{2s_1} + \frac{2s_2}{(s_1^2-s_2^2)} & \frac{s_2}{s_1(s_1-s_2)} \\ \hline \frac{s_1}{s_2(s_2-s_1)} & -\frac{1}{2s_2} + \frac{2s_1}{(s_2^2-s_1^2)} \end{array}\right]$$

By definition

$$B_{ij}^n = \sum_{k=1}^{n} \frac{4s_k s_j T_k^n T_j^n}{(s_i+s_k)^2(s_k+s_j)} \tag{3.22}$$

We now establish an inductive principle to derive B_{ij}^n along lines very similar to that used for Δ_{ij}^n. As in equations (3.14) to (3.16) it is readily shown that

$$B_{ij}^n - B_{ij}^{n-1} = \left[\frac{2s_j T_j^n}{(s_n+s_j)T_n^n}\right] B_{in}^n. \tag{3.23}$$

Now make the inductive hypothesis

$$B_{ij}^{n-1} = \frac{s_j T_j^{n-1}}{s_i(s_i-s_j)T_i^{n-1}} = \frac{s_j(s_n-s_j)(s_n+s_i)T_j^n}{s_i(s_i-s_j)(s_n+s_j)(s_n-s_i)T_i^n} \quad i \neq j. \tag{3.24}$$

Using (3.23) and (3.24)

$$B_{ij}^n - \left[\frac{s_j T_j^n}{s_i(s_i-s_j)T_i^n}\right] = B_{ij}^{n-1} + \left[\frac{2s_j T_j^n}{(s_n+s_j)T_n^n}\right] B_{in}^n - \left[\frac{s_j T_j^n}{s_i(s_i-s_j)T_i^n}\right]$$

$$= \frac{-s_j T_j^n}{s_i(s_i-s_j)T_i^n}\left[1 - \frac{(s_n-s_j)(s_n+s_i)}{(s_n+s_j)(s_n-s_i)}\right] + \frac{2s_j T_j^n B_{in}^n}{(s_n+s_j)T_n^n}$$

or

$$\left[B^n_{ij} - \frac{s_j T^n_j}{s_i(s_i-s_j)T^n_i}\right] - \frac{2s_jT^n_j}{(s_n+s_j)T^n_n}\left[B^n_{in} - \frac{s_nT^n_n}{s_i(s_i-s_n)T^n_i}\right] = 0 \tag{3.25}$$

or

$$B_{ij} = \frac{s_jT_j}{s_i(s_i-s_j)T_i} - \frac{2s_jT_j}{(s_n+s_j)T_n}\left[B_{in} - \frac{s_nT_n}{s_i(s_i-s_n)T_i}\right].$$

As in Theorem 2, from the symmetry in (3.22) it is necessary that

$$B_{ij} = \frac{s_jT_j}{s_i(s_i-s_j)T_i} \qquad i \neq j \tag{3.26}$$

and (3.21a) is proved by induction. When $i=j$,

$$B^{n-1}_{ii} = -4s_iT^{n-1}_i\sum_{k=1}^{n-1}\frac{s_kT^{n-1}_k}{(s_i+s_k)^3} = -4s_i\frac{(s_n-s_i)}{(s_n+s_i)}T^n_i\sum_{k=1}^{n-1}\frac{s_kT^n_k(s_n-s_k)}{(s_i+s_k)^3(s_n+s_k)}$$

$$= -4s_iT^n_i\frac{(s_n-s_i)}{(s_n+s_i)}\sum_{k=1}^{n}\frac{s_kT^n_k(s_n-s_k)}{(s_i+s_k)^3(s_n+s_k)} \tag{3.27}$$

Then

$$B^n_{ii} - B^{n-1}_{ii} = -4s_iT^n_i\sum_{k=1}^{n}\frac{s_kT^n_k}{(s_i+s_k)^3}\left[\frac{(s_n-s_i)(s_n-s_k)}{(s_n+s_i)(s_n+s_k)} - 1\right]$$

$$= -\frac{2s_iT^n_i}{(s_n+s_i)T^n_n}B^n_{in}. \tag{3.28}$$

From (3.26) and (3.28)

$$B^n_{ii} - B^{n-1}_{ii} = \frac{2s_n}{s_i^2 - s_n^2} \tag{3.29}$$

and (3.21b) is proved by induction.

Q.E.D.

III.1.2.3 Equations of the Direct Method in Integral Form

An explicit expression has now been found for the matrix $\underline{\underline{B}}$ in equations (3.12) and the equations of Aigrain and Williams have been reduced to a set of n equations and n unknowns in the $\{s_k\}$ alone. Although equations (3.12) appear in the simplest form possible, they are highly nonlinear in the $\{s_k\}$ and serious computational difficulties arise in their solution. With a little manipulation, these equations may be expressed in an integral form which, as we shall see in the next section, has an illuminating geometrical interpretation. Even more important, this form is well suited for solution using a linear iterative method.

By multiplying each equation in (3.12) by $s_i T_i^n$ one gets

$$s_i T_i^n F'(-s_i) - \sum_{k=1}^{n} s_i T_i^n B_{ik}^n F(-s_k) = 0 \quad i=1,2,\ldots,n. \qquad (3.30)$$

It can immediately be shown that (3.30) can be written in the more compact form

$$\int_{-j\infty}^{j\infty} \frac{H(s)}{s-s_i} \frac{F(-s)ds}{2\pi j} = 0 \qquad i=1,2,\ldots,n \qquad (3.31)$$

where

$$H(s) = \prod_{k=1}^{n} \frac{(s+s_k)}{(s-s_k)}. \qquad (3.32)$$

This is easily verified by using residue calculus and noting that

$$\frac{\partial T_i^n}{\partial s_i} = -\left(B_{ii}^n + \frac{1}{2s_i}\right) T_i^n. \qquad (3.33)$$

The relation of equation (3.31) to the Kautz procedure for construction of orthogonal functions and associated development by others, is discussed next.

III.1.3 Method of McDonough and Huggins

In a recent paper [7] McDonough and Huggins suppressed the amplitude coefficients $\{\alpha_k\}$ in the Aigrain-Williams equations by regarding f(t) as a signal in a vector space. Their argument proceeded as follows: Let the error of the approximation be $e(t)=f(t)-f_a(t)$. Then it is readily seen that equations (3.2a) and (3.2b) may be written respectively as

$$\left.\begin{aligned} &\int_0^\infty e(t)\exp(s_k t)\,dt = 0 \\ \text{and}\quad &\\ &\int_0^\infty e(t)\,t\exp(s_k t)\,dt = 0 \end{aligned}\right\}\ k=1,2,\dots,n. \tag{3.34}$$

In the language of vector spaces, equation (3.34) means that the error e(t) is orthogonal to the space S_{2n} which is spanned by the 2n "vectors" $\{\exp(s_i t),\ t\exp(s_i t)\}$. Also, by definition, the approximating function $f_a(t)$ must lie in the subspace S_n spanned by the n vectors $\{\exp(s_i t)\}$. Let $S_{2n}-S_n$ denote the subspace of S_{2n} that is complementary to S_n and let $\{\phi_{n+i}(t)\}$ i=1,2,...,n be basis for this subspace. A basis for $S_{2n}-S_n$ can be formed by applying the Gram-Schmidt procedure to the functions $(\exp(s_1 t),\dots,\exp(s_n t),\ t\exp(s_1 t),\dots,t\exp(s_n t))$ taken in that order, to construct the orthonormal basis functions $\{\phi_i\}$, i=1,2,...,2n. Clearly, $\{\phi_{n+i}\}$ i=1,2,...,n is orthogonal to both $f_a(t)$ and e(t) and hence, must be orthogonal to $f(t) = e(t) + f_a(t)$. The $\Phi_{n+i}(s)$, which are the Laplace transforms of $\phi_{n+i}(t)$, may be constructed by simple inspection from Kautz's method. That is,

$$\Phi_{n+i}(s) = H(s)\,\Phi_i(s) \tag{3.35}$$

where again

$$H(s) = \prod_{k=1}^{n} \frac{(s+s_k)}{(s-s_k)} \tag{3.36}$$

and

$$\phi_i(s) = \frac{\sqrt{-s_i-s_i^*}}{s-s_i} \prod_{k=1}^{i-1} \frac{(s+s_k^*)}{(s-s_k)} . \tag{3.37}$$

Now, n independent equations of constraint on the $\{s_k\}$ must be

$$\int_0^\infty f(t)\ \phi_{n+i}(t)\ dt = 0 \qquad i=1,2,\dots,n, \tag{3.38}$$

which when written in the frequency domain, using the Parseval relation, are

$$\int_{-j\infty}^{j\infty} F(-s)\ \phi_{n+i}(s)\ H(s)\ \frac{ds}{2\pi j} = 0 \qquad i=1,2,\dots,n. \tag{3.39}$$

Since these equations involve the $\{s_k\}$ only, the $\{\alpha_k\}$ have been suppressed as in (3.31). In fact, it will be shown in the next section that the left-hand sides of equations (3.39) are merely linear combinations of the left-hand sides of equations (3.31). These equations, (3.31) or (3.39), are still nonlinear in terms of the unknowns $\{s_k\}$ and apparently one of the best ways to solve for them is by a numerical linear iterative scheme first suggested by Sears [32] and described as follows:

Let the all-pass[†] operator H(s) in (3.39) (or (3.31)) be replaced by the more general operator

$$H_a(s) = \sum_{k=1}^{n+1} \frac{b_k s^{k-1}}{D(s)} = \frac{(-1)^n D_a(-s)}{D(s)} \quad , b_{n+1} = 1 \tag{3.40}$$

where

$$D(s) = \prod_{k=1}^{n} (s-s_k). \tag{3.41}$$

[†] H(s) is sometimes called the all-pass operators for if s is replaced by $j\omega$, the magnitude of H(s) is one, independent of the value of the frequency ω.

If $H_a(s)$ is used in (3.39) instead of $H(s)$, one obtains n simultaneous equations which, being linear in the $\{b_1,b_2,\ldots,b_n\}$, may be written in matrix form as

$$\underline{\underline{M}}\ \underline{B} = \underline{Z} \tag{3.42}$$

where $\underline{\underline{M}}$ is an n X n matrix with i,k^{th} element

$$m_{ik} = \int_{-j\infty}^{j\infty} [F(s)\ s^{k-1}/D(s)]\ \Phi_i(s)\ \frac{ds}{2\pi j}\ . \tag{3.43}$$

$\underline{B}$ and $\underline{Z}$ are columns with elements

$$z_i = -\ m_{i,n+1} \tag{3.44}$$

and b_i are the unknown coefficients. The iterative algorithm is as follows:

(a) Given the poles at the j^{th} iteration, i.e.,

$\{s_1,s_2,\ldots,s_n\}_j$

evaluation the matrix $(\underline{\underline{M}})_j$ and the vector $(\underline{Z})_j$.

(b) Solve equation (3.42), to obtain the coefficients of the vector $(\underline{B})_{j+1}$

(c) From $(\underline{B})_{j+1}$ find the new pole locations $\{s_1,\ldots,s_n\}_{j+1}$ using $(D_a(s))_{j+1} = 0$.

(d) Repeat from (a) with j=j+1. Continue the process until the change $\max_i |(s_i)_j-(s_i)_{j+1}|$ is less than some small pre-assigned value.

The convergence properties of this method shall be discussed in chapter IV.

III.1.4 Equivalence of the Direct Method and McDonough's Method

Let
$$\Psi_i(s) = H(s)\ \frac{1}{s-s_i} \qquad i=1,2,\ldots,n. \tag{3.45}$$

Then from Kautz's method it is easily seen that the $\{\Psi_i(s)\}$ also form a basis for the difference space $S_{2n}-S_n$ and hence, equations (3.31) have the same geometric interpretation used by McDonough. Thus,

$$\int_0^\infty F(-s)\ \Psi_i(s)\ \frac{ds}{2\pi j} = 0 \qquad i=1,2,\ldots,n \tag{3.46}$$

are just linear combinations of equations (3.39). Notice that the all-pass function H(s) is still preserved and the same linear iterative scheme can be used.

These new equations in terms of $\Psi_i(s)$ have two advantages over (3.39). First, $\Psi_i(s)$ has only one double pole whereas $\Phi_{n+i}(s)$ has i double poles. This means the old set will have i-1 extra derivative terms when the residues are evaluated and, moreover, each term will have (i-1) extra factors of the form $(s_k+s_i)/(s_k-s_i)$. Clearly, there is a saving in computational time by using the new equations. Second, equation (3.46) may be written in the matrix form

$$\underline{F}' = \underline{\underline{B}}\ \underline{F}. \tag{3.47}$$

This enables one to use matrix algebra to find the $\{s_k\}$ using (3.21). Solution of the equation in this form has not been attempted here, but is a topic for further investigation.

While the $\Phi_{n+i}(s)$ are orthonormal, the $\Psi_i(s)$ are not. As a result,[†]

[†] One is contrasting an <u>orthogonal</u> versus an <u>oblique</u> basis, both of which span the same space. Naturally, there will always be more correlation between the oblique elements. However, here, as is usually the case, the oblique elements are easier to express mathematically and any gains in accuracy made by using orthonormal elements may be lost due to the extra complexity introduced into these expressions.

although equations (3.46) have the simpler form

$$\sum_{k=1}^{n}\left[\int_{-j\infty}^{j\infty} F(-s)\,\frac{s^{k-1}}{D(s)}\,\frac{1}{(s-s_i)}\,\frac{ds}{2\pi j}\right] b_k = -\int_{-j\infty}^{j\infty} F(-s)\,\frac{s^n}{D(s)(s-s_i)}\,\frac{ds}{2\pi j}$$

$$= \sum_{k=1}^{n} p_{ik} b_k = r_i \qquad i=1,2,\ldots,n$$

or

$$\underline{\underline{P}}\;\underline{B} = \underline{R}, \tag{3.48}$$

they are "softer" than equations (3.42).

Evaluating the elements of the matrix $\underline{\underline{P}}$ by residue calculus, one gets

$$p_{ik} = \frac{(-s_i)^{k-1}\left[F'(-s_k) + \left\{(k-1) + s_i \sum_{\substack{m=1\\ m\neq i}}^{n} \frac{1}{(s_i-s_m)}\right\} F(-s_i)\right]}{\Gamma_i}$$

$$+ \sum_{\substack{m=1\\ m\neq i}}^{n} \frac{(-s_m)^{k-1}\, F(-s_m)}{(s_m-s_i)\,\Gamma_m} \tag{3.49}$$

where

$$\Gamma_i = \prod_{\substack{m=1\\ m\neq i}}^{n} (s_i-s_m) \tag{3.50}$$

Since f(t) is real, the $\{s_k\}$ must occur in complex conjugate pairs. Upon examining (3.49) it is seen that if s_i is replaced by s_i^*, p_{ik} becomes p_{ik}^*. This also implies that if s_i is real, so is p_{ik}. Hence, it ir possible to avoid complex arithmetic altogether in finding the real $\{b_i\}$ from (3.48) by using the equivalent system of n equations

$$\sum_{j=1}^{n} p_{ij}\, b_j = r_i \qquad i=1,2,\ldots,\text{NREAL} \tag{3.51a}$$

$$\left.\begin{array}{l} \mathrm{Re}\ \{ \sum_{j=1}^{n} p_{ij} b_j = r_i \} \\ \\ \mathrm{Im}\ \{ \sum_{j=1}^{n} p_{ij} b_j = r_i \} \end{array}\right\} \begin{array}{l} i= \text{NREAL} +1,\dots,n-1 \text{ in} \\ \quad \text{steps of } 2 \end{array} \tag{3.51b}$$

where NREAL is the number of real roots. If an i corresponding to a complex valued s_i is used in (3.51b), an i corresponding to s_i^* will give the identical equations and so should be omitted.

III.1.5 Method of McBride, Schaefgen, and Steiglitz

In this section we examine the method of finding the approximation of f(t) by exponentials due to McBride, Schaefgen, and Steiglitz [6] (hereafter referred to as the MSS method). They start with the approximating function expressed in the frequency domain as

$$F_a(s) = \frac{a_1+a_2s+\dots+a_n s^{n-1}}{b_1+b_2s+\dots+b_n s^{n-1}+s_n} = \frac{N(s)}{D(s)} \tag{3.52}$$

instead of the equivalent partial fraction expansion. Then, the functional

$$J = \int_0^\infty [f(t)-f_a(t)]^2\, dt = \int_0^\infty e^2(t)\, dt$$

is to be minimized over the 2n real coefficients $\{a_k,b_k\}$. Necessary conditions at the minimum are that

$$\left.\begin{array}{l} \frac{\partial J}{\partial a_k} = 0 = 2 \int_0^\infty e(t) \frac{\partial e(t)}{\partial a_k}\, dt \\ \\ \frac{\partial J}{\partial b_k} = 0 = 2 \int_0^\infty e(t) \frac{\partial e(t)}{\partial b_k}\, dt \end{array}\right\} \quad k=1,2,\dots,n \tag{3.53}$$

Equations (3.53) are nonlinear in the $\{a_k,b_k\}$ and one is faced with the same difficulties in solving them as with the equations of Aigrain and Williams. The key feature of the MSS method is the introduction of an approximate error $E_a(s)$, viz.

$$E_a(s) = \frac{D_j(s)}{D_{j-1}(s)} F(s) - \frac{N_j(s)}{D_{j-1}(s)} \qquad (3.54)$$

where now the subscript j refers to the iteration number. To solve the equations in a feasible way, the previously computed $(b_k)_{j-1}$ coefficients of D_{j-1} are regarded as fixed during the j^{th} iteration. This linearizes the error in terms of the unknown coefficients $\{a_k, b_k\}_j$ of the numerator polynomials N_j and D_j. One now replaces e(t) in (3.53) by $e_a(t)$, the inverse transform of $E_a(s)$, and uses an iterative scheme very similar to the one employed by McDonough described earlier. With repeated iterations $D_{j-1}(s)$ approaches $D_j(s)$ and thus $E_a(s)$ approaches the true error $E(s)=F(s)-F_a(s)$.

However, inserting $e_a(t)$ in (3.53) has three distinct disadvantages. First, instead of utilizing the convenient point form of the Aigrain-Williams equations, one must evaluate a set of 2n partial derivatives and then integrate. The resulting equations are much more complicated than (3.4). Second, and more important, the Mode-1 Iteration used on these 2n new equations does not converge to the optimum solution since the approximate error is minimized rather than the true error. Thus, following the Mode-1 Iteration, a Mode-2 Iteration is also needed to further refine the results and find the optimum solution. This difficulty does not arise in the Mode-2 Iteration because the expressions are correct in the limit as the $\{b_k\}$ approach their optimum values. Third, several examples will demonstrate that Mode-2 converges more slowly than McDonough's method and a new method which we will present in section (III.1.6). These two Modes will now be examined in detail.

Mode-1 Iteration of the MSS Method

To minimize the functional

$$J_a = \int_0^\infty e_a(t)\, e_a(t)\, dt \tag{3.55}$$

over the 2n coefficients $\{a_k, b_k\}$, one requires that J_a be stationary with respect to changes in the parameters,

$$\left.\begin{aligned} \frac{\partial J_a}{\partial a_i} &= 0 = 2\int_0^\infty e_a(t)\, \frac{\partial e_a(t)}{\partial a_i} = 0 \\ \frac{\partial J_a}{\partial b_i} &= 0 = 2\int_0^\infty e_a(t)\, \frac{\partial e_a(t)}{\partial b_i} = 0 \end{aligned}\right\} \; i=1,2,\ldots,n. \tag{3.56}$$

Observe that

$$\frac{\partial E_a(s)}{\partial a_i} = -\frac{s^{i-1}}{D_j(s)}$$

and

$$\frac{\partial E_a(s)}{\partial b_i} = \frac{s^{i-1}F(s)}{D_j(s)} \qquad i=1,2,\ldots,n. \tag{3.57}$$

Using the Parseval relation on (3.56) one gets

$$\frac{1}{2}\frac{\partial J_a}{\partial a_i} = \int_{-j\infty}^{j\infty} \left[\frac{D_j(-s)F(-s)-N_j(-s)}{D_{j-1}(-s)}\right]\left[\frac{s^{i-1}}{D_{j-1}(s)}\right]\frac{ds}{2\pi j} = 0 \tag{3.58a}$$

$$\frac{1}{2}\frac{\partial J_a}{\partial b_i} = \int_{-j\infty}^{j\infty} \left[\frac{D_j(-s)F(-s)-N_j(-s)}{D_{j-1}(-s)}\right]\left[\frac{s^{i-1}F(s)}{D_j(s)}\right]\frac{ds}{2\pi j} = 0 \tag{3.58b}$$

$$i=1,2,\ldots,n.$$

Hence, equations (3.58) provide another linear iterative scheme involving 2n real parameters $\{a_k, b_k\}$ instead of just the n $\{b_k\}$. Otherwise, the iterations are carried out as in the McDonough method.

Mode-2 Iteration of the MSS Method

The error E(s) may be written

$$E(s) = F(s) - F_a(s) = F(s) - \frac{N(s)}{D(s)} . \qquad (3.59)$$

Thus

$$\frac{\partial E(s)}{\partial a_i} = \frac{-s^{i-1}}{D(s)}$$

$$\frac{\partial E(s)}{\partial b_i} = \frac{-s^{i-1}N(s)}{D^2(s)} = \frac{-s^{i-1}}{D(s)} F_a(s) \qquad (3.60)$$

Using the Parseval relation on (3.53) gives the conditions for the stationarity of the integrated square error in the frequency domain as

$$\int_{-j\infty}^{j\infty} \left[F(-s) - \frac{N(-s)}{D(-s)}\right] \left[\frac{s^{i-1}}{D(s)}\right] \frac{ds}{2\pi j} = 0 \qquad (3.61a)$$

$$\int_{-j\infty}^{j\infty} \left[F(-s) - \frac{N(-s)}{D(-s)}\right] \left[\frac{s^{i-1}}{D(s)} \frac{N(s)}{D(s)}\right] \frac{ds}{2\pi j} = 0 \qquad (3.61b)$$

$$i=1,2,\ldots,n$$

If the iterative process for Mode-1 converges, $D_{j-1}(s)$ approaches $D_j(s)$ and comparison with (3.61a) shows that (3.58a) is correct in the limit. However, (3.58b) is not correct in the limit which is observed when it is compared with (3.61b). For this reason, Mode-1 Iteration does not converge in general to the optimum solution. This difficulty may be eliminated by using (3.60) to change (3.58b) to

$$\int_{-j\infty}^{j\infty} \left[\frac{D_j(-s)F(-s)-N_j(-s)}{D_{j-1}(s)}\right] \left[\frac{s^{i-1}}{D_{j-1}(s)} \frac{N_{j-1}(s)}{D_{j-1}(s)}\right] \frac{ds}{2\pi j} = 0 \qquad (3.62)$$

Equations (3.58a) and (3.62) are now used in the Mode-2 Iteration. Convergence to the optimum solution is now usually possible, but as we shall see in chapter IV, Mode-2 converges so slowly that in order to make the MSS method practical, one must first use the more rapidly converging Mode-1 Iteration to bring one "near enough" to the optimum point in parameter space. Furthermore, there is no guarantee Mode 2 converges if one is not "near enough" since the equations that determine it are not correct unless one is actually at the minimum.

III.1.6 The New Method

Thus far, we have discussed two linear iterative schemes in sections III.1.3 and III.1.5. Each has worked well for the cases reported and appears to be useful in finding by numerical computation the matched exponents for the approximation of a known time function. This section develops yet another linear iterative method which offers the advantages of both the methods described in sections III.1.3 and III.1.5 and reveals the link between them. This method leads to the same results as those described by McDonough and Huggins.

Fundamental Equations

The Aigrain-Williams equations (3.4) may be written in the form.

$$\left.\begin{aligned} E(-s_k) = F(-s_k) - F_a(-s_k) = 0 \\ E'(-s_k) = F'(-s_k) - F_a'(-s_k) = 0 \end{aligned}\right\} \quad k=1,2,\dots,n. \qquad (3.63)$$

The equations in this form suggest that a better way of using the approximate error $E_a(s)$ defined by (3.54) is to impose the constraints of equations (3.63) directly upon it. This immediately yields a set of new linear equations for the j^{th} iteration.

$$\left.\begin{aligned} E_a(s) = 0 \quad (3.64a) \\ E_a'(s) = 0 \quad (3.64b) \end{aligned}\right\} \text{for } s = (-s_k)_{j-1} \qquad k=1,2,\dots,n$$

where the $(s_k)_{j-1}$ are the roots of the denominator D_{j-1} obtained from the previous iteration. Equations (3.64a) may be expressed as

$$D_jF-N_j = 0, \quad s = (-s_k)_{j-1} \qquad k=1,2,\ldots,n \qquad (3.65a)$$

Similarly, upon differentiating $E_a(s)$ with respect to s, equations (3.64b) may be written as

$$E' = \{D_{j-1}(D_jF'+D_j'F) - D_{j-1}'D_jF+(D_{j-1}'N_j-N_j'D_{j-1})\}/D_{j-1}^2$$

$$= 0, \quad s = (-s_k)_{j-1} \qquad k=1,2,\ldots,n$$

which by using (3.65a), simplifies to

$$F'D_j-F\,D_j' = N_j', \qquad s = (-s_k)_{j-1} \qquad k=1,2,\ldots,n \qquad (3.65b)$$

The 2n simultaneous equations (3.65) are linear in terms of the unknowns $\{a_k,b_k\}$. The iterative procedure is carried out in the same way as described in the two previous methods using equations (3.65) or the equivalent equations (3.67) for convenient computation to find the $\{a_k,b_k\}_j$. The initial point in parameter space may be determined perhaps from Prony's method or Padé approximants.

A very important consequence of imposing these constraints upon $E_a(s)$ is that the D_{j-1} required in the formulation of the MSS method does not appear in equations (3.65); the introduction of the approximate error $E_a(s)$ was unnecessary. In fact, an appropriate set of linear equations may be gotten directly from the Aigrain-Williams constraints as follows:

From (3.4a) and (3.52)

$$N(s) = D(s)\,F(s) \quad \text{for} \quad s = (-s_i) \qquad i=1,2,\ldots,n \qquad (3.66a)$$

where the $\{s_i\}$ are the roots of the n^{th} degree polynomial D(s). Also,

$$F' = (-D'\,N + N'\,D)/D^2$$

From (3.68), (3.69) and Theorem 4, $\underline{\underline{C}}$ can be written explicitly in terms of the $\{s_i\}$ as

$$c_{ik} = h_{ik} = \sum_{j=1}^{n} q_{ij}(-s_i)^{j-1} F(-s_i) \qquad (3.70)$$

and $\underline{U}$ may be regarded as the negative of the n+1th column of $\underline{\underline{C}}$ or

$$u_i = - c_{i,n+1} \qquad i=1,2,\ldots,n. \qquad (3.71)$$

But (3.48), (3.49), and (3.70) reveal that

$$\Gamma_i \, p_{ij} = c_{ij}$$

and thus McDonough's method and the one developed here must be equivalent.

Discussion

In this section we have revealed the strong link between the MSS method and that of McDonough. Both methods use equation (3.54) or its equivalent to linearize the iterative process (although this was not so obvious in the latter). The crucial difference in the methods is that the MSS method considers variation of the error with respect to the $\{a_k, b_k\}$ parameters, whereas in McDonough's method (and the one developed here), the variation with respect to the exponents $\{s_k\}$ (and the hidden $\{\alpha_k\}$) is considered. It is not possible to write a set of linear equations in the $\{a_k, b_k\}$ for the true error surface whereas equations (3.65) and (3.66) show that one may do this when the variation is with respect to the $\{\alpha_k, s_k\}$, thus avoiding the need for two types of iterations.

In conclusion, the new method developed here has several computational advantages over the one developed by MSS. First, it requires only one iterative scheme instead of two. Second, by using the point form of the Aigrain-Williams equations, the matrices in equation (3.67) immediately appear as explicit function of the $\{a_1\}$. In the MSS method the corresponding matrix elements (see Table 4.1, p. 61) are much harder to evaluate. Third, as we shall show in chapter IV, for all examples thus far examined, the new method converges more quickly to the matched exponents than the MSS method.

IV. CONVERGENCE AND COMPARISON OF THE LINEAR ITERATIVE SCHEMES

In the previous chapter three linear iterative schemes were described. Two of these, the method of McDonough and the new method, were shown to be equivalent in their results, although computationally different! That is, for any initial $\{s_i\}$ and fixed n, the resulting iterations of either of these methods will be identical barring roundoff errors. However, because the new method uses the rational fraction form of $F_a(s)$, a direct comparison of it, rather than McDonough's scheme, with the MSS method will be made since this will be much easier to do.

IV.1 Comparison of the Iterative Equations

The 2n equations used in the iterative scheme of MSS may be written in matrix form as

$$\begin{aligned} \underline{\underline{V1}}\,\underline{A} + \underline{\underline{G1}}\,\underline{B} &= \underline{X1} \\ \underline{\underline{W1}}\,\underline{A} + \underline{\underline{H1}}\,\underline{B} &= \underline{Y1} \end{aligned} \qquad (4.1)$$

for Mode-1 Iteration and as

$$\begin{aligned} \underline{\underline{V1}}\,\underline{A} + \underline{\underline{G1}}\,\underline{B} &= \underline{X1} \\ \underline{\underline{W2}}\,\underline{A} + \underline{\underline{H2}}\,\underline{B} &= \underline{Y2} \end{aligned} \qquad (4.2)$$

for Mode-2 Iteration. From (3.58a) it is seen that the elements of $\underline{\underline{V1}}$ are

$$\begin{aligned} v1_{ik} &= -\int_{-j\infty}^{j\infty} \frac{(-s)^{k-1}\, s^{i-1}}{D_{j-1}(-s)\, D_{j-1}(s)} \frac{ds}{2\pi j} \\ &= \frac{(-1)^{n-k}}{2} \sum_{k=1}^{n} \left[\frac{(s_k)^{i+k-3}}{\prod\limits_{\substack{m=1 \\ m\neq k}}^{n} (s_k{}^2 - s_m{}^2)} \right] \qquad i,k=1,2,\ldots,n, \end{aligned} \qquad (4.3a)$$

compared with the corresponding much simpler expression $v_{ik} = (-s_i)^{k-1}$ given in equation (3.67) for the new method, and

$$g1_{ik} = \int_{-j\infty}^{j\infty} \frac{(-s)^{i-1}\, s^{k-1} F(s)}{D_{j-1}(-s)\, D_{j-1}(s)} \frac{ds}{2\pi j} \tag{4.3b}$$

$$x1_i = -g1_{i,n+1}. \tag{4.3c}$$

From (3.58b) it is shown that the remaining elements in equation (4.1) are

$$w1_{ik} = -\int_{-j\infty}^{j\infty} \frac{(-s)^{k-1}\, s^{i-1} F(s)}{D_{j-1}(-s)\, D_{j-1}(s)} \frac{ds}{2\pi j} \tag{4.4a}$$

$$h1_{ik} = -\int_{-j\infty}^{j\infty} \frac{(-s)^{i-1}\, s^{k-1} F(s)\, F(-s)}{D_{j-1}(s)\, D_{j-1}(-s)} \frac{ds}{2\pi j} \tag{4.4b}$$

and

$$y1_i = -h_{i,n+1}. \tag{4.4c}$$

Equation (4.4b) can be difficult to evaluate by residue calculus. For example, if f(t) is the square pulse, $F(s) = (1-e^{-s})/s$, the product of F(s) F(-s) with any rational function of s will have an essential singularity at infinity, in both the right and left hand planes, and thus, direct evaluation of (4.4b) by residues requires special treatment.[†] (Of course, one may evaluate these Fourier transforms by direct integration with respect to ω over $-\infty$ to ∞ without resorting to residues but this is usually extremely difficult.) Alternatively, one may invoke the Parseval relation and evaluate the equivalent time-domain equations (3.56) to obtain the matrix elements for Mode-1 Iteration.

Table 4.1 summarizes all of these equations and clearly shows that the elements in the new method are much easier to compute.

[†] See Appendix C for details.

Table 4.1 - Comparison of the Methods

New Method	Mode-1 Iteration	Mode-2 Iteration
$v_{ik} = (-s_i)^{k-1}$	$v1_{ik} = (-1)^k \int_{-j\infty}^{j\infty} \frac{s^{i+k-2}}{D_{j-1}(s)\, D_{j-1}(-s)} \frac{ds}{2\pi j}$ $= \frac{(-1)^{n-k}}{2} \sum_{m=1}^{n} \left[\frac{(s_m)^{i+k-3}}{\prod_{\substack{\ell=1 \\ \ell \neq m}}^{n} (s_m^2 - s_\ell^2)} \right]$	$v2_{ik}$ - Same as Mode-1
$g_{ik} = -(-s_i)^{k-1}\, F(-s_i)$	$g1_{ik} = (-1)^{k-1} \int_{-j\infty}^{j\infty} \frac{s^{i+k-2}\, F(-s)}{D_{j-1}(s)\, D_{j-1}(-s)} \frac{ds}{2\pi j}$	$g2_{ik}$ - Same as Mode-1
$x_i = -\, g_{i,n+1}$	$x1_i = -\, g1_{i,n+1}$	$x2_i = -\, g1_{i,n+1}$
$w_{ik} = (k-1)\, (-s_i)^{k-2}$	$w1_{ik} = (-1)^i \int_{-j\infty}^{j\infty} \frac{s^{i+k-2}\, F(-s)}{D_{j-1}(s)\, D_{j-1}(-s)} \frac{ds}{2\pi j}$	$w2_{ik} = (-1)^j \int_{-j\infty}^{j\infty} \frac{s^{i+k-2}\, N_{j-1}(s)}{D_{j-1}(-s)\, D_{j-1}^2(s)} \frac{ds}{2\pi j}$
$h_{ik} = (-s_i)^{k-1}\, F'(-s_i)$ $+(k-1)(-s_i)^{k-2} F(-s_i)$	$h1_{ik} = (-1)^{i-1} \int_{-j\infty}^{j\infty} \frac{s^{i+k-2} F(s) F(-s)}{D_{j-1}(s) D_{j-1}(-s)} \frac{ds}{2\pi j}$	$h2_{ik} = (-1)^{k-1} \int_{-j\infty}^{j\infty} \frac{s^{i+k-2} F(-s) N_{j-1}(s)}{D_{j-1}(-s)\, D_{j-1}^2(s)} \frac{ds}{2\pi j}$
$y_i = -\, h_{i,n+1}$	$y1_i = -\, h1_{i,n+1}$	$y2_i = -\, h2_{i,n+1}$

IV.2 Rates of Convergence

To predict by mathematical analysis the rate and region of convergence of any of the linear iterative schemes for a general f(t) is extremely difficult except for the simple case when n=1. Instead, we provide several numerical examples to give the reader some feel for results obtained by the different methods. For any f(t) that is composed exactly of n exponentials, any of the iterative processes, Mode-1, Mode-2, and the new method, will yield these exponentials immediately after one iteration. Consequently, when f(t) is "nearly" exponential, one would also expect reasonably rapid convergence for any of the methods. This is indeed the case as we now illustrate by specific examples.

Numerical Results - Consider the two time functions

$$f_1(t) = (e^{-t} + e^{-2t})\, u_{-1}(t) \tag{4.5}$$

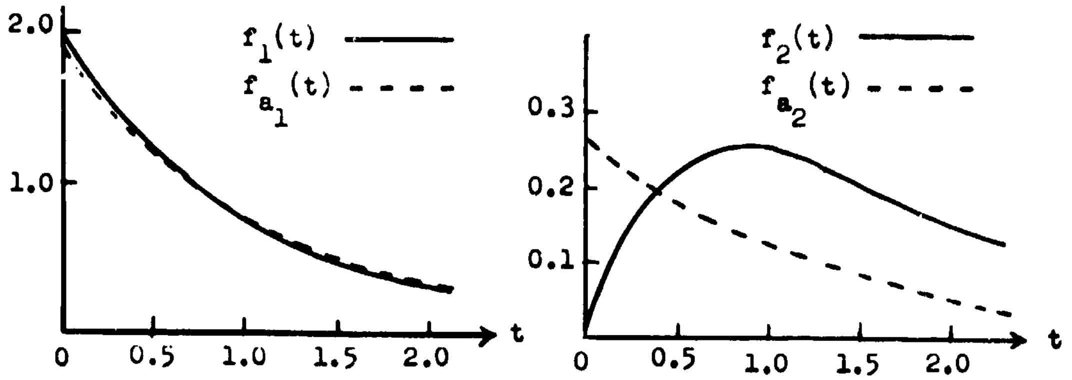
and

$$f_2(t) = (e^{-t} - e^{-2t})\, u_{-1}(t) \tag{4.6}$$

each to be approximated by a single exponential. Figure 1 shows qualitatively why $f_1(t)$ can be approximated very accurately by one exponential whereas $f_2(t)$ cannot.

Table 4.2 gives the result for $f_1(t)$ of the iterations by the various methods all starting with an initial value of $s_1 = -1.2$. The new method and Mode-2 Iteration converge to the same result (the optimum approximation) but Mode-2 required 500 iterations whereas the new method needed only 4. In contrast, Mode-1 converged as rapidly as the new method, but not, to the optimum approximation. Thus, for this simple signal, the new method is superior to both Mode-1 and Mode-2.

Figure 1 -- Plots of $f_1(t)$ and $f_2(t)$ and the best approximations of them by a single exponential.



****

For $f_2(t)$, shown in Table 4.3, the Mode-2 Iteration took about 3000 cycles! Also, observe that the Mode-1 error is considerably larger than in the case for $f_1(t)$. This is reasonable since $f_1(t)$ more "nearly" resembles a single exponential than does $f_2(t)$. (Recall that Mode-1 only gives exact values when $f(t)$ is an exponential.)

From equation (3.54), the linearized error for approximating the square pulse, $F(s)=(1-e^{-s})/s$, by a single exponential becomes

$$e_a(t) = \frac{(b_1)_j}{(b_1)_{j-1}} f_3(t)-(a_1)_j\, e^{-(b_1)_{j-1}t} - \frac{[(b_1)_j-(b_1)_{j-1}]}{(b_1)_{j-1}}\, e^{-(b_1)_{j-1}t}\,\{u_{-1}(t)-e^{-(b_1)_{j-1}}\,u_{-1}(t-1)\} \qquad (4.7)$$

where

$$f_3(t) = u_{-1}(t)-u_{-1}(t-1).$$

Table 4.4 reveals that Mode-2 Iteration and the new method converge to the optimum exponent at the same rate. (By coincidence, the iterations are almost identical for this case. They differ in 8th or 9th decimal place.)

Table 4.2 - Approximation of the function $f_1(t)=[\exp(-t)+\exp(-2t)]u_{-1}(t)$ by One Exponential.

j	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$
		1.20000		1.20000		1.20000
1	1.84197	1.20139	1.93499	1.32265	1.93369	1.32095
2	1.84309	1.20277	1.93779	1.32639	1.93908	1.32815
3	1.8442	1.20416	1.93787	1.3265	1.93938	1.32856
4	1.84531	1.20555	1.93788	1.3265	1.9394	1.32859
5			1.93788	1.3265	1.9394	1.32859
⋮	⋮	⋮				
100	1.92123	1.30392				
101	1.9216	1.30442				
102	1.92197	1.30492				
103	1.92233	1.30541				
104	1.92269	1.30588				
⋮	⋮	⋮				
200	1.93739	1.32584				
201	1.93744	1.3259				
202	1.93748	1.32596				
203	1.93753	1.32602				
204	1.93757	1.32608				
⋮	⋮	⋮				
300	1.93919	1.32831				
301	1.9392	1.32831				
302	1.9392	1.32832				
303	1.93921	1.32833				
304	1.93921	1.32833				
⋮	⋮	⋮				
400	1.93938	1.32856				
401	1.93938	1.32856				
402	1.93938	1.32856				
403	1.93938	1.32856				
⋮	⋮	⋮				
500	1.9394	1.32859				
	Mode-2 Iteration		Mode-1 Iteration		New Method	

****

Table 4.3 - Approximation of the function $f_1(t)=[\exp(-t)-\exp(-2t)]u_{-1}(t)$ by One Exponential.

j	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$
		5.00000		5.00000		5.00000
1	.233677	4.81444	.121212	.090909	.076923	-1.76923
2	.238613	4.63995	.267525	.519313	-1.85714	2.0989
3	.243427	4.47615	.244642	.417085	.138931	-.334183
4	.248103	4.32267	.249851	.43871	.428824	.809871
5			.248736	.434005	.216462	.290949
6			.248978	.435023	.279182	.53473
7			.248926	.434802	.245733	.421202
8			.248937	.43485	.260254	.474336
9			.248935	.43484	.25325	.44952
10			.248935	.434842	.256473	.461122
11			.248935	.434841	.254956	.4557
12			.248935	.434841	.255663	.458234
13					.255332	.45705
14					.255487	.457603
15					.255414	.457345
16					.255488	.457466
17					.255432	.457409
18					.25544	.457436
19					.255436	.457423
20					.255438	.457429
21					.255437	.457426
22					.255437	.457428
23					.255437	.457427
24					.255437	.457427
⋮	⋮	⋮				
100	.329639	2.1215				
101	.329687	2.12001				
102	.329735	2.11855				
⋮	⋮	⋮				
300	.332367	2.03381				
301	.332371	2.03369				
302	.332375	2.03357				
⋮	⋮	⋮				
3000	.255437	.457427				
	Mode-2 Iteration		Mode-1 Iteration		New Method	

In this example the term containing the factor $[(b_1)_j-(b_1)_{j-1}]$ has little affect on the equations that determine the iterations. However, it seems quite possible that the extraneous terms, which always arise when using Mode-1 or Mode-2, that contain the factor $[(b_1)_j-(b_1)_{j-1}]$ could sometimes affect the equations enough to prevent convergence of Mode-2 if $(b_1)_j$ is not "near enough" to its optimum value.

Table 4.4 - Approximation of the Square Pulse by One Exponential

j	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$	$(a_1)_j$	$(b_1)_j$
		1.00000		1.00000		1.00000
1	1.51217	1.39221	1.38344	1.18857	1.51217	1.39221
2	1.39272	1.188	1.3638	1.14261	1.39272	1.188
3	1.4511	1.29183	1.36851	1.15348	1.4511	1.29183
4	1.42045	1.23836	1.36739	1.15089	1.42045	1.23836
5	1.43597	1.26572	1.36766	1.15151	1.43597	1.26572
6	1.42796	1.25167	1.36759	1.15136	1.42796	1.25167
7	1.43206	1.25887	1.36761	1.1514	1.43206	1.25887
8	1.42995	1.25518	1.3676	1.15139	1.42995	1.25518
9	1.43103	1.25707	1.36761	1.15139	1.43103	1.25707
10	1.43048	1.2561	1.36761	1.15139	1.43048	1.2561
11	1.43076	1.2566			1.43076	1.2566
12	1.43061	1.25634			1.43061	1.25634
13	1.43069	1.25648			1.43069	1.25648
14	1.43063	1.25641			1.43065	1.25641
15	1.43067	1.25644			1.43067	1.25644
16	1.43066	1.25643			1.43066	1.25643
17	1.43067	1.25643			1.43067	1.25643
18	1.43066	1.25643			1.43066	1.25643
19	1.43066	1.25643			1.43066	1.25643
	Mode-2 Iteration		Mode-1 Iteration		New Method	

****

Table 4.5 shows the results fitting a square pulse using 3 exponentials with the initial values of the parameters chosen as (s_i) = (-1,-2,-3). After 90 iterations the new method had converged to s_1= -2.246602, $s_{2,3}$= -1.443643 ±j4.150741 which is in agreement to 6 significant figures with McDonough's result, [1] p. 159, found by a search method.

Table 4.5 - Three Exponential Approximation of the Square Pulse.

Iteration	Exponents After That Iteration	
	-3.0	-2.0, -1.0
1	-3.047813	-2.037829 ± j3.433864
2	-2.065591	-1.153318 ± j3.881929
3	-2.540835	-1.786042 ± j4.094648
4	-2.050865	-1.185975 ± j4.093363
5	-2.437354	-1.686921 ± j4.160736
6	-2.091982	-1.246309 ± j4.125865
7	-2.385592	-1.625349 ± j4.165192
8	-2.129335	-1.296223 ± j4.134689
9	-2.350106	-1.580867 ± j4.162651
10	-2.158346	-1.334265 ± j4.139171
11	-2.324005	-1.547741 ± j4.159904
12	-2.180300	-1.362868 ± j4.142197
13	-2.304550	-1.522938 ± j4.157694
14	-2.196827	-1.384329 ± j4.144391
.	.	.
.	.	.
.	.	.
45	-2.247180	-1.449385 ± j4.150813
46	-2.246105	-1.448001 ± j4.150679
47	-2.247036	-1.449200 ± j4.150795
48	-2.246230	-1.448162 ± j4.150694
49	-2.246927	-1.449060 ± j4.150781
50	-2.246323	-1.448283 ± j4.150706
.	.	.
.	.	.
.	.	.
82	-2.246600	-1.448639 ± j4.150741
83	-2.246607	-1.448647 ± j4.150742
84	-2.246602	-1.448640 ± j4.150741

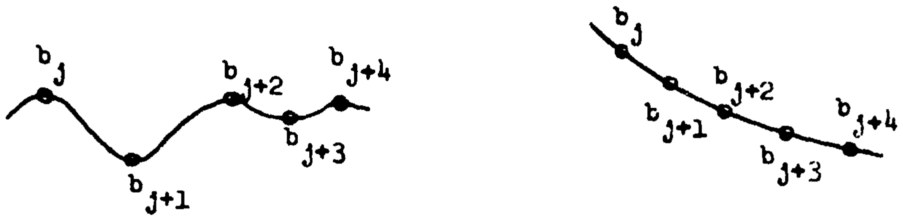
****

IV.3 Accelerated Convergence - Shanks' Method

Although the new method converged to the optimum solution faster then the MSS method (even assuming a switch from Mode-1 to Mode-2 is made at the best time, a decision which is apparently ad hoc) in every case tested, it too converged rather slowly in some cases - 90 iterations with n=3 for the square pulse and 24 iterations with n=1 for $f_2(t)$. For larger n it seems that convergence would be e en slover. In an attempt to speed up convergence one may incorpcrate a little used method due to Shanks [35].

Consider any of the numerical sequences in Tables 4.2 through 4.5 which are either monotonic or oscillatory. Draw a smooth curve through these discrete points. Typical graphs are depicted in figure 2.

Figure 2 -- Graphs demonstrating transient characteristics in the iterative sequences.



New Method -- $f_2(t)$, $f_3(t)$
Mode-1 -- $f_2(t)$, $f_3(t)$
Mode-2 -- $f_3(t)$

New Method -- $f_1(t)$
Mode-1 -- $f_1(t)$
Mode-2 -- $f_1(t)$, $f_2(t)$

****

These graphs look like first- or second-order transients. By "n^{th} order transient" we mean any function which has the form

$$p(t) = B + \sum_{i=1}^{n} c_i \exp(-\alpha_i t) \qquad \text{Re } \{\alpha_i\} > 0.$$

Shanks' method predicts the limit B of such sequences by "filtering out" or annihilating the exponential components.

Tables 4.6 shows the result of applying Shanks' method to 2 sequences obtained by using the new method. In both cases the thirteenth iteration is correct to only two decimal digits or so, but e_5 is already correct to six digits. In any event, extreme caution must be exercised when applying Shanks' method to these sequences since there is no theory to justify its use here. However, it has been demonstrated how helpful the method can sometimes be in reducing the number of iterations needed and is a topic worthy of further investigation.

Table 4.6a - Shanks' Method Applied to the First 13 Iterations of the Matched Exponent of a One Exponential Approximation of $f_2(t)$ by the New Method.

j	$(b_1)_j$	e_1	e_2	e_3	e_4	e_5
1	5.0000000D 00	6.4230761D-01	8.2149931D-01	4.5629824D-01	4.5747099D-01	4.5743029D-01
2	1.7692300D 00	6.0530278D-01	4.5232274D-01	4.5725943D-01	4.5742968D-01	4.5743017D-01
3	2.0989000D 00	4.4397785D-01	4.6007478D-01	4.5742277D-01	4.5743036D-01	4.5742709D-01
4	-3.3418000D-01	4.5287860D-01	4.5733626D-01	4.5742996D-01	4.5743020D-01	
5	8.0987400D-01	4.5681384D-01	4.5744401D-01	4.5743033D-01	4.5743171D-01	
6	2.9095200D-01	4.5727632D-01	4.5742927D-01	4.5743017D-01		
7	5.3473300D-01	4.5739919D-01	4.5743046D-01	4.5742967D-01		
8	4.2120500D-01	4.5742337D-01	4.5743009D-01			
9	4.7433900D-01	4.5742885D-01	4.5742992D-01			
10	4.4952300D-01	4.5742986D-01				
11	4.6112500D-01	4.5742992D-01				
12	4.5570300D-01					
13	4.5823700D-01					
	• • •					
24	4.5742700D-01					

****

Table 4.6b - Shanks' Method Applied to the First 13 Iterations of the Real Matched Exponent of a Three Exponential Approximation of $f_3(t)$ Using the New Method.

j	$(-s_1)_j$	e_1	e_2	e_3	e_4	e_5
1	3.0478130D 00	2.3858696D 00	2.2564312D 00	2.2482896D 00	2.2471412D 00	2.2465814D 00
2	2.0655910D 00	2.2995877D 00	2.2462530D 00	2.2477584D 00	2.2467970D 00	2.2466069D 00
3	2.5408350D 00	2.2669253D 00	2.2485199D 00	2.2473050D 00	2.2466160D 00	2.2466097D 00
4	2.0508650D 00	2.2549663D 00	2.2471252D 00	2.2472023D 00	2.2466072D 00	
5	2.4373540D 00	2.2506792D 00	2.2473202D 00	2.2468652D 00	2.2466106D 00	
6	2.0919820D 00	2.2487596D 00	2.2469435D 00	2.2470340D 00		
7	2.3855920D 00	2.2479321D 00	2.2458741D 00	2.2466626D 00		
8	2.1293350D 00	2.2474833D 00	2.2467206D 00			
9	2.3501060D 00	2.2472242D 00	2.2466747D 00			
10	2.1583460D 00	2.2470535D 00				
11	2.3240050D 00	2.2469356D 00				
12	2.1803000D 00					
13	2.3045500D 00					
	• • •					
85	2.2466070D 00					

****

V. DISCUSSION OF RESULTS AND AREAS FOR FURTHER WORK

Throughout this work we have assumed the real function f(t) to be: known analytically for all time; piecewise continuous; and of bounded energy, $\int_0^\infty f^2(t)\,dt < \infty$. Under these three restrictions we reviewed in chapter III several ways of finding a linear combination of n exponentials to yield a least-squares approximation of the function over the semi-infinite interval. The results presented in chapter IV show the new method is the best of these for finding the matched exponents. This method requires that the function F(s) and its first derivative be evaluated only at the n points $s= - s_i$ in the right-half of the s-plane. In the vicinity of these points the function is always well-behaved, as may be seen from the Cauchy-Schwartz inequality,

$$|F(-s_i)|^2=|\int_0^\infty f(t)\exp(s_i t)\,dt|^2 \le [\int_0^\infty f^2(t)dt][\int_0^\infty |\exp(s_i t)|^2\,dt]$$

or

$$|F(-s_i| \le \frac{1}{\sqrt{2Re\{-s_i\}}}[\int_0^\infty f^2(t)dt]^{1/2} \tag{5.1}$$

provided $(-s_i)$ is in the right half plane.

The restriction that the signal be expressed initially as an analytic function of time can also be removed provided the signal is expressed on some other basis such as $f= \sum_k c_k \phi_k$ for which the $\Phi_k(s)$ are known in the right half plane. The next section gives an important example in which the signal is represented initially on a discrete primal basis.

V.1 The New Method Applied to Sampled Data

Let $p(t) = \sum_{k=-\infty}^{\infty} \delta(t-kT)$ denote an impulse train with the impulses spaced T second apart. The sampling of a function can be described mathematically by multiplication with p(t). That is

$$f^*(t)=p(t)f(t)=\sum_{k=+\infty}^{\infty} f(t)\delta(t-kT)=\sum_{k=-\infty}^{\infty} f(kT)\delta(t-kT) \tag{5.2}$$

It is easily shown [33] that the Laplace transform of f*(t) is

$$F^*(s)=\int_0^{\infty} f^*(t)e^{-st}\,dt=\sum_{k=-\infty}^{\infty} f(kt)e^{-kTs} \tag{5.3}$$

Consider the change in variable z=exp(Ts) which maps the left half plane in the s domain inside the unit circle in the z domain. Then the Z-transform of the function f(t) (f(t)=0 t < 0) is defined to be

$$F(z)=\sum_{k=0}^{\infty} f(kT)z^{-k}. \tag{5.4}$$

The approximating function at these sampled instants, is given by the rational Z-transform:

$$F_a(z)=\frac{N(z)}{D(z)}=\frac{a_1+a_2\,z^{-1}+\ldots+a_n\,z^{-(n-1)}}{1+b_1\,z^{-1}+\ldots+b_n z^{-n}}$$

$$=\frac{\alpha_1}{\frac{1}{z}-z_1}+\frac{\alpha_2}{\frac{1}{z}-z_2}+\ldots+\frac{\alpha_n}{\frac{1}{z}-z_n} \tag{5.5}$$

The poles of $F_a(z)$ must all be inside the unit circle to ensure stability and thus $|z_k| > 1$ k=1,2,...n. The error at these sampled instants is defined to be

$$e(kT) = f(kT) - f_a(kT) \tag{5.6}$$

and

$$E(z) = F(z) - F_a(z). \tag{5.7}$$

Then (Ragazzini p. 179)

$$J=\sum_{k=0}^{\infty}[e(kT)]^2=\frac{1}{2\pi j}\oint_{\substack{\text{unit}\\\text{circle}}} E(z)E(\tfrac{1}{z})\,\frac{dz}{z} \tag{5.8}$$

The necessary conditions on the 2n parameters $\{\alpha_i, z_i\}$ to minimize the functional J are

$$\frac{\partial J}{\partial \alpha_i} = 0 = 2\frac{1}{2\pi j}\oint \frac{E(z)\,dz}{z(z-z_i)} \tag{5.9a}$$

$$\frac{\partial J}{\partial z_k} = 0 = 2\frac{1}{2\pi j}\oint \frac{E(z)\,dz}{z(z-z_i)^2} \tag{5.9b}$$

From the Cauchy integral formula and the fact that E(z) has all its poles inside the unit circle

$$\begin{aligned} &E(z_i) = 0 \\ &E'(z_i) = 0 \qquad i=1,2,\ldots,n \end{aligned} \tag{5.10}$$

Equations (5.10) are intuitively correct since they are the Aigrain-Williams equations applied to sampled data. Notice $|z_i| > 1$ corresponds to a point in the right half plane in the frequency domain. These equations are solved iteratively exactly as before except one uses F(z) instead of F(s).

Steiglitz and McBride [34] have also applied their more complicated method to sampled data.

V.2 Concluding Remarks

For large n (n>5) double-precision arithmetic is required to get meaningful results using the new method. This is not unexpected since the same difficulty arises in the simpler linear least-squares approximation discussed in chapter II. Based on experience with the two methods, it is of the author's opinion that the roundoff errors in this nonlinear approximation will be about the same order of magnitude as those in chapter II. The computational aspects of this method deserve additional study, but they will not be pursued further in this thesis because they involve considerations foreign to the main thrust of this work.

APPENDIX A

Construction of Orthonormal Functions

Equation (2.11) can also be used in a reverse manner so that if $\underline{\underline{G}}^{-1}$ is known, one can sometimes construct an orthonormal basis by simple inspection and avoid the Gram-Schmidt method altogether. With Gastinel's result (derived without regard to orthogonal functions) and use of (2.11) it is possible to derive Kautz's important result for orthogonalizing exponentials. This second application is demonstrated by the following example.

A General Formula for Orthonormal Polynomials with Respect to a Constant Weight Function

Define

$$x_i = t^{s_i - 1/2}, \quad 0 \leq t \leq 1,\ s_i \geq 1/2,$$

then

$$g_{ij} = \int_0^1 t^{s_i - 1/2}\, t^{s_j - 1/2}\, dt = \frac{1}{(s_i + s_j)} \quad i,j = 1,2,\ldots,n. \qquad (A.1)$$

As shown previously, the inverse of this n X n symmetric matrix is

$$g_{ij}^{-1} = \frac{4 s_i s_j}{(s_i + s_j)} T_i \, T_j \qquad (A.2)$$

where

$$T_m = \prod_{\substack{k=1 \\ k \neq m}}^{n} \frac{s_k + s_m}{s_k - s_m} \, .$$

From (2.17) and (A.2)

$$c_{nn}^2 = g_{nn}^{-1} = 2 s_n T_n^2 . \qquad (A.3)$$

Also $c_{nn} c_{nj} = g_{nj}^{-1}$, and from (A.2) and (A.3)

$$c_{nj} = (2 s_n)^{1/2} \frac{2 s_j}{(s_n + s_j)} T_j . \qquad (A.4)$$

Since the formula must hold for any n,

$$c_{ij}=(2s_i)^{1/2}\frac{2s_j}{(s_i+s_j)}\prod_{\substack{k=1\\k\neq j}}^{i}\frac{s_k+s_j}{s_k-s_j}\qquad j\leq i. \tag{A.5}$$

Hence by (2.3), (2.4) and (A.5)

$$\phi_i(t)=\sum_{j=1}^{i}c_{ij}t^{s_j-1/2}\quad . \tag{A.6}$$

However, this set is orthonormal on (0,1). To generalize to (a,b), consider the linear transformation t=kt' + d. When t=0,t' = a and when t=1, t' = b. So k=1/(b-a) and d= -a/(b-a). Hence, the general formula is[†]

$$\phi_i(t')=(b-a)^{-1/2}\sum_{j=1}^{n}\left\{\frac{(2s_i)^{1/2}(2s_j)}{(s_i+s_j)}T_j\left[\frac{(t'-a)}{(b-a)}\right]^{s_j-1/2}\right\} \tag{A.7}$$

Note that there is no requirement $s_j-1/2$ be an integer. Now let $P_n(t)$ denote the Legendre polynomial of degree n. It is not hard to show from (A.7) that, in this special case for which $s_j-1/2 = j-1$,

$$P_n(t)=(-1)^n\sum_{j=1}^{n}\frac{(n-j-2)!(t+1)^{j-1}}{(-2)^{j-1}(n-j)!((j-1)!)^2} \tag{A.8}$$

where[††]

$$\int_{-1}^{1}P_n(t)P_m(t)dt=\left(\frac{1}{2n+1}\right)\delta_{nm}. \tag{A.9}$$

[†] Note that $\int_0^1\phi_i(t)\ \phi_j(t)\ dt=\int_a^b\phi_i(t')\ \phi_i(t')\ dt'/(b-a)=\delta_{ij}$. Hence, the factor $(b-a)^{-1/2}$ appears in (A.7).

[††] The factor $(-1)^n$ can be dropped and the polynomials will still be orthonormal. This factor is added to make (A.8) agree with the standard Legendre polynomials.

The Determinant of the Gram Matrix

Let D_n be the determinant of the n X n matrix $\underline{\underline{G}}$. Then

$$\det(\underline{\underline{G}}^{-1})=\det(\tilde{\underline{\underline{C}}}\underline{\underline{C}})=\det(\tilde{\underline{\underline{C}}})\det(\underline{\underline{C}})=1/D_n. \qquad (A.10)$$

But $\underline{\underline{C}}$ is a triangular matrix and its determinant is just the product of its diagonal terms. Hence

$$D_n=\left(\prod_{k=1}^{n} c_{kk}^2\right)^{-1}. \qquad (A.11)$$

Also, it is seen that†

$$c_{nn}=\left(\frac{D_{n-1}}{D_n}\right)^{1/2} . \qquad (A.12)$$

For the Hilbert matrix

$$c_{nn}=(-1)^{n+1}(2s_n)^{1/2}\prod_{k=1}^{n-1}\frac{(s_k+s_n)}{(s_k-s_n)}, \qquad (A.13)$$

or

$$D_n(\text{Hilbert})=\frac{\prod\limits_{i<j}\left[\frac{s_i-s_j}{s_i+s_j}\right]^2}{\prod\limits_{k=1}^{n}(2s_k)} \qquad (A.14)$$

$i=1,2,\ldots,n=1 \quad j=2,3,\ldots,n.$

Similarly, for the matrix discussed with the Laguerre basis considered in the Appendix,

$$D_n(\text{Laguerre})=\left\{\prod_{k=1}^{n}(k-1)!\right\}^2 , \qquad (A.15)$$

since $c_{nn}=(-1)^{n-1}/(n-1)!$.

† Szego, [25] sec. 11.1.10, recognized this formula for orthonormal polynomials.

Concluding Remarks

The method described at the beginning of chapter II shows a way of finding a closed form inverse of some Gram matrices that often occur in linear least-squares problems, provided an analytic expression for an appropriate set of orthonormal functions can be found in terms of the original basis elements. If an analytic expression cannot be found for the orthonormal functions, the Gram-Schmidt procedure can always be used. But then the method loses some of its merit, for if the basis elements are highly correlated, one may encounter the new difficulty of computing the elements of $\underline{\underline{C}}$ accurately.

Another distinct advantage of this method over the "direct" use of orthonormal functions in least-squares is that it will reveal common factors that may be present in each term of the inverse. This is illustrated by equation (B.5) in Appendix B which shows the common factor $[(i-1)!(j-1)!]^{-2}$ of each term in the inverse of $\underline{\underline{G}}$(Laguerre). It is unlikely that this common factor would have been observed if linear combinations of the orthonormal functions were used to reconstruct the original basis. Finding such factors when they exist can obviously save time and improve computational accuracy. The results for the Hilbert matrix are even better.

Perhaps more important than the direct application to the least-squares problem is the possibility of constructing orthonormal bases by simple inspection from $c_{nj}=g_{nj}^{-1}/(g_{nn}^{-1})^{1/2}$ when an explicit expression for g_{ij}^{-1} can be found (as in the case of the Hilbert matrix). The construction of the orthonormal basis for fractional powers of t was achieved by this method.

APPENDIX B

A Least-Squares Problem Using Laguerre Polynomials

The Laguerre polynomial $L_n(t)$ is a polynomial of degree n in t for which

$$\int_0^\infty e^{-t}\, L_n(t)L_m(t)\, dt = \delta_{nm}. \tag{B.1}$$

Laguerre polynomials are orthonormal with respect to the weight function e^{-t} over $(0, \infty)$. It is known that

$$L_{n-1}(t) = \sum_{k=1}^{n} \binom{n-1}{k-1} \frac{(-t)^{k-1}}{(k-1)!} \tag{B.2}$$

so that

$$\phi_n(t) = e^{-t/2} L_{n-1}(t). \tag{B.3}$$

Suppose one desires to find the a_k such that a continuous function $f(t)$ is approximated in the least-square sense over $(0, \infty)$ by

$$f_a(t) = \sum_{k=1}^{n} a_k(t^{k-1}e^{-t/2}) = \sum_{k=1}^{n} a_k x_k(t).$$

(This may appear to be an odd choice of the x_k, but they are much easier to work with than $e^{-t/2}L_n(t)$, just as integrals involving single terms of the form $\exp(s_k t)$ are easier to evaluate analytically than integrals involving orthogonal functions formed from the exponentials.) Then, as in (2.3) it is immediately seen from (B.2) that

$$\begin{aligned} c_{ij} &= \binom{i-1}{j-1} \frac{(-1)^{j-1}}{(j-1)!} = \frac{(i-1)!(-1)^{j-1}}{(i-j)!((j-1)!)^2} & i \geq j \\ &= 0 & i < j. \end{aligned} \tag{B.4}$$

From (2.11) one finds[†]

[†] At first glance one might think that $g_{ij}^{-1} = c_{ni}c_{nj}$, as in the Hilbert matrix. However, the result held there regardless of the ordering of the s_i. Here (B.4) holds only for a particular ordering of the basis elements and the generalisation cannot be made. Hence, the summation is necessary.

$$g^{-1}_{ij} = \sum_{k=i}^{n} c_{ki} c_{kj}$$

$$= \sum_{k=i}^{n} \binom{k-1}{i-1} \frac{(-1)^{i-1}}{(i-1)!} \binom{k-1}{j-1} \frac{(-1)^{j-1}}{(j-1)!} \qquad (B.5)$$

$$= \frac{(-1)^{i+j}}{((i-1)!(j-1)!)^2} \sum_{k=i}^{n} \frac{((k-1)!)^2}{(k-i)!(k-j)!}$$

where

$$g_{ij} = \langle x_i , x_j \rangle = \int_0^\infty e^{-t} \, t^{i+j-2} \, dt = (i+j-2)! \qquad (B.6)$$

and

$$f_j = \int_0^\infty f(t) t^{j-1} e^{-1/2} \, dt \qquad j=1,2,\ldots,n. \qquad (B.7)$$

Hence, the solution in closed form is $\underline{\underline{A}} = \underline{\underline{G}}^{-1}\underline{\underline{F}}$. (Assuming that (B.7) may be evaluated in closed form.)

APPENDIX C

Numerical Example for Mode-1 Iteration

Consider the function

$$f_4(t) = e^{-\alpha t}\,[u_{-1}(t) - u_{-1}(t-1)]$$

to be approximated by one exponential. The Laplace transform of $f_4(t)$ is

$$F_4(s) = \frac{1-e^{-(s+\alpha)}}{s+\alpha}$$

and

$$E_a(s) = \frac{s+(b_1)_j}{s+(b_1)_{j-1}}\,F_4(s) - \frac{(a_1)_j}{s+(b_1)_{j-1}}$$

From Table 4, page 61

$$\begin{bmatrix} v_{11} & g_{11} \\ w_{11} & h_{11} \end{bmatrix} \begin{bmatrix} (a_1)_j \\ (b_1)_j \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

where

$$v_{11} = -\int_{-j\infty}^{j\infty} \frac{1}{[-s+(b_1)_{j-1}][s+(b_1)_{j-1}]}\,\frac{ds}{2\pi j} = -\frac{1}{2(b_1)_{j-1}}$$

$$g_{11} = \int_{-j\infty}^{j\infty} \frac{F_4(s)}{[s+(b_1)_{j-1}][-s+(b_1)_{j-1}]}\,\frac{ds}{2\pi j} = \frac{F_4((b_1)_{j-1})}{2(b_1)_{j-1}}$$

$$x_1 = -\int_{-j\infty}^{j\infty} \frac{sF_4(s)}{(s+(b_1)_{j-1})(-s+(b_1)_{j-1})}\,\frac{ds}{2\pi j} = -\frac{F_4((b_1)_{j-1})}{2}$$

$$w_{11} = -\int_{-j\infty}^{j\infty} \frac{F_4(s)}{(s+(b_1)_{j-1})(-s+(b_1)_{j-1})}\,\frac{ds}{2\pi j} = -\frac{F_4((b_1)_{j-1})}{2(b_1)_{j-1}}$$

and

$$h_{11} = \int_{-j\infty}^{j\infty} \frac{F_4(s)F_4(-s)}{(s+(b_1)_{j-1})(-s+(b_1)_{j-1})}\,\frac{ds}{2\pi j}\,.$$

To evaluate the last integral by residues one must be particularely careful because the product $F_4(s)F_4(-s)$ has an essential singularity at $s=\infty$.

Hence, one cannot make use of Jordan's lemma, [36] p. 300, to directly evaluate the integral by residues. However, the integral may be broken down into the sum of two parts, one which vanishes along the infinite semi-circular arc containing the left-half plane and the other which vanishes along the infinite arc containing the right-half plane. That is

$$h_{11} = \underset{\text{RHP}}{\int_{-j\infty}^{j\infty}} \frac{1+e^{-2\alpha}-e^{-(s+\alpha)}}{(s+\alpha)(s+(b_1)_{j-1})(-s+\alpha)(-s+(b_1)_{j-1})} \frac{ds}{2\pi j}$$

$$+ \underset{\text{LHP}}{\int_{-j\infty}^{j\infty}} \frac{-e^{-(\alpha-s)}}{(s+\alpha)(s+(b_1)_{j-1})(-s+\alpha)(-s+(b_1)_{j-1})} \frac{ds}{2\pi j}$$

which simplifies to

$$h_{11} = \frac{1+e^{-2\alpha}-2e^{-(\alpha+(b_1)_{j-1})}}{2(b_1)_{j-1}(\alpha^2-(b_1)^2_{j-1})} + \frac{1-e^{-2\alpha}}{2\alpha((b_1)^2_{j-1}-\alpha^2)} .$$

Finally

$$y_1 = - h_{12} = -\int_{-j\infty}^{j\infty} \frac{sF_4(s)F_4(-s)}{D_{j-1}(s)\; D_{j-1}(-s)} \frac{ds}{2\pi j}$$

$$= \frac{1+e^{-2\alpha}}{2(\alpha^2-(b_1)_{j-1})} + \frac{1+e^{-2\alpha}}{2((b_1)^2_{j-1}-\alpha^2)} = 0.$$

An iterative algorithm very similar to the one described on page 46 is then used to find $(a_1)_j$ and $(b_1)_j$.

REFERENCES

[1] R. N. McDonough, "Representation and analysis of signals; Pt. 15, matched exponents for the representation of signals", Dept. of Elec. Engrg., The Johns Hopkins University, Baltimore, Md., DDC Doc. AD 411431, April 30, 1963.

[2] W. H. Huggins, "The use of orthogonalized exponentials", Report AFCRC TR-57-357, AD 133741, 15 Nov. 1958.

[3] T. Y. Young and W. H. Huggins, "On the representation of electrocardiograms", IEEE Trans. BME-10, No. 3, pp. 86-95, July 1963.

[4] T. Y. Young and W. H. Huggins, " 'Complementary' signals and orthogonalized exponentials", IRE Trans. CT-9, No. 4, Dec. 1962.

[5] R. E. A. C. Paley and N. Wiener, Fourier Transforms in the Complex Domain, Am. Math. Soc. Colloqu. Pubs., Vol. XIX, Am. Math. Soc., N. Y., 1934.

[6] L. E. McBride, Jr., H. W. Schaefgen, and K. Steiglitz, "Time-domain approximation by iterative methods", IEEE Trans. CT-13, No. 4, pp. 381-387, Dec. 1966.

[7] R. N. McDonough and W. H. Huggins, "Best least-squares representation of signals by exponentials", IEEE Trans. AC-13, No. 4, pp. 408-412, August 1968.

[8] D. T. Tang, "The Tchebysheff approximation of a prescribed impulse response with RC network realization", 1961 IRE International Convention Record, Part 4, pp. 214-220.

[9] W. H. Kautz, "Network synthesis for a specified transient response", Tech. Rep. No. 209, Res. Lab. of Electronics M.I.T., Cambridge, Mass., 1952.

[10] G. C. W. Mathers, "Synthesis of lumped-circuits for optimum transient response", Electronics Research Lab., Stanford University, Stanford, California, Nov. 1951.

[11] R. D. Teasdale, "Time domain approximation by use of Padé approximants", 1953 Convention Record of the I.R.E. Part 5 - Circuit Theory pp. 210-216.

[12] P. R. Aigrain and E. M. Williams, "Synthesis of n-reactive networks for desired transient response", Journal of Applied Physics, Vol. 20 pp. 597-600, June 1949.

References (cont'd.)

[13] W. H. Kautz, "Transient synthesis in the time domain", *IRE Trans.* CT-1, No. 3, pp. 29-39, Sept. 1954.

[14] N. Gastinel, "Inversion d'une matrice generalisant le matrice de Hilbert", *Chiffres* Vol. 3, pp. 149-152, Sept. 1960.

[15] F. E. Hildebrand, *Introduction to Numerical Analysis* , McGraw-Hill, 1956, pp. 429-439.

[16] N. Newman and J. Todd, "On the evaluation of matrix inversion programs", *J. Soc. Indust. Appl. Math.*,6, pp. 466-476, 1958.

[17] J. von Neumann and H. H. Goldstine, "Numerical inverting of matrices of high order", *Bull. Am. Math. Soc.*, 53, pp. 1021-1099, 1947.

[18] W. H. Huggins, "Signal theory", *IRE Trans. Circuit Theory*, Vol. CT-3, pp. 210-216, December 1956.

[19] G. Golub, "Numerical methods for solving linear least squares problems", *Numerische Math.*, Vol. 7, pp. 206-216, 1965.

[20] J. R. Rice, *The Approximation of Functions*, Reading, Mass.: Addison-Wesley, 1964, pp. 30-47.

[21] J. Todd, "Computational problems concerning the Hilbert matrix", *J. Research NBS*, Vol. 65B pp. 19-22, January, 1961.

[22] I. R. Savage and E. Lukacs, "Tables of inverses of finite segments of the Hilbert matrix", *NBS Appl. Math. Ser.*, 39, pp. 105-108, 1954.

[23] A. R. Collar, "On the reciprocal of a segment of a generalized Hilbert matrix", *Proc. Cambridge Phil. Soc.*, Vol. 47, pp. 11-17, 1951.

[24] R. B. Smith, "Two theorems on inverses of finite segments of the generalized Hilbert matrix", *Math. Tables Aids Comput.*, Vol. 13, pp. 41-43, 1959.

[25] G. Szëgo, *Orthonormal Polynomials*, rev. ed. New York: Am. Math. Soc. Colloq. Publ. 1959, Vol. 23.

[26] G. Pólya and G. Szëgo, *Aufgaben und Lehrsatz aus der Analysis*, Vol. 2, 3rd Corrected printing (Grundlehren der Mathenschaften Wissematischen Ser., Vol. 20) Berlin: Springer, 1964, p.98.

[27] J. T. Tou, *An Introduction to Modern Control Theory* , McGraw-Hill, New York, N. Y.; pp. 51-53, 1964.

References (cont'd.)

[28] J. D. Brulé, "A note on the Vandermonde determinant", IEEE Trans. on Automatic Control (Correspondence), Vol. AC-9, pp. 314-315, July 1964.

[29] J. L. Tou, "Determination of the inverse Vandermonde matrix", IEEE Trans. on Automatic Control (Correspondence), Vol. AC-9, p. 314, July 1964.

[30] K. Hoffman and R. Kunze, Linear Algebra , Prentice Hall, Inc., Englewood Cliffs, N.J.; 1961.

[31] G. E. P. Box, "The exploration and exploitation of response surfaces: some general considerations and examples", Biometrics, Vol. 10, No. 1, pp. 16-60, 1954.

[32] R. W. Sears, Jr., "Digital optimization of exponential representations of signals", Ph.D. dissertation, Dept. of Elec. Engrg., Johns Hopkins University, Baltimore, Md., 1966.

[33] J. R. Ragazzini and G. F. Franklin, Sampled-Data Control Systems, New York: McGraw-Hill, 1958.

[34] K. Steiglitz and L. E. McBride, "A technique for identification of linear systems", IEEE Trans. Vol. AC-10, No. 4, Oct., 1965, pp. 461-464.

[35] D. Shanks, "Non-linear transformations of divergent and slowly convergent sequences", Journal of Mathematics and Physics Vol.34 , pp. 1-42, 1955.

[36] A. Papoulis, The Fourier Integral and its Application, New York: McGraw-Hill, 1962.